JA999 254

日　本　国　特　許　庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９９年１２月２７日

出願番号
Application Number: 平成１１年特許願第３７１３４７号

出願人
Applicant (s): インターナショナル・ビジネス・マシーンズ・コーポレーション

２０００年　６月２９日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

出証番号　出証特２０００－３０５１３０７

【書類名】　特許願

【整理番号】　JA999254

【提出日】　平成11年12月27日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F　3/00

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

【氏名】　土方　嘉徳

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

【氏名】　青木　義則

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

【氏名】　中島　周

【特許出願人】

【識別番号】　390009531

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレイション

【代理人】

【識別番号】　100086243

【弁理士】

【氏名又は名称】　坂口　博

【復代理人】

【識別番号】　100104880

【弁理士】

【氏名又は名称】　古部　次郎

【選任した代理人】

【識別番号】　100091568

【弁理士】

【氏名又は名称】　市位　嘉宏

【選任した復代理人】

【識別番号】　100100077

【弁理士】

【氏名又は名称】　大場　充

【手数料の表示】

【予納台帳番号】　081504

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9304391

【包括委任状番号】　9304392

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　情報抽出システム、情報処理装置、情報収集装置、文字列抽出方法及び記憶媒体

【特許請求の範囲】

【請求項１】　通信ネットワークにて接続されたサーバとクライアントとを備えた情報抽出システムであって、

前記サーバは、

前記クライアントにおいて閲覧に供されるデータファイルを提供し、

前記クライアントは、

前記通信ネットワークを介して前記サーバから受信した前記データファイルの内容を表示する閲覧手段と、

前記閲覧手段にて表示された前記データファイルの内容を閲覧する際にユーザが行った入力操作に基づいて予め定められた特定の操作を検出する操作検出手段と、

前記操作検出手段により検出された、前記閲覧手段の表示画面上における前記特定の操作の行われた場所に表示されている情報を抽出する手段とを備えることを特徴とする情報抽出システム。

【請求項２】　ウェブコンテンツを格納するウェブサーバと、通信ネットワークを介して当該ウェブサーバからウェブコンテンツを受信して表示するクライアントとを備えた情報抽出システムであって、

前記クライアントは、前記ウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、

前記ウェブサーバに格納された前記ウェブコンテンツは、前記クライアントの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された当該操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、前記ウェブコンテンツの中から当該特定の操作の対象となった情報を抽出して前記ウェブサーバに返送する処理とを前記クライアントに実行させる、前記クライアントの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを埋め込んであることを特徴とする情報抽出システム。

【請求項３】　ウェブコンテンツを格納するウェブサーバと、通信ネットワークを介して当該ウェブサーバからウェブコンテンツを受信して表示するクライアントとを備えた情報抽出システムであって、

前記クライアントは、前記ウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、

前記ウェブサーバは、前記クライアントの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された当該操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、前記ウェブコンテンツの中から当該特定の操作の対象となった情報を抽出して前記ウェブサーバに返送する処理とを前記クライアントに実行させる、前記クライアントの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを前記ウェブコンテンツに埋め込んで前記クライアントに送信することを特徴とする情報抽出システム。

【請求項４】　ウェブコンテンツを格納するウェブサーバと、通信ネットワークを介して当該ウェブサーバからウェブコンテンツを受信して付加的な処理を行うプロキシサーバと、当該プロキシサーバにて当該付加的な処理を施された当該ウェブコンテンツを表示するクライアントとを備えた情報抽出システムであって、

前記クライアントは、前記ウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、

前記プロキシサーバは、前記クライアントの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された当該操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、前記ウェブコンテンツの中から当該特定の操作の対象となった情報を抽出して前記プロキシサーバに返送する処理とを前記クライアントに実行させる、前記クライアントの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを、前記ウェブサーバから受信した前記ウェブコンテンツに埋め込んで前記クライアントに送信することを特徴とする情報抽出システム。

【請求項５】　ウェブコンテンツを格納するウェブサイトと、通信ネットワークを介して当該ウェブサイトからウェブコンテンツを受信して表示するウェブ

ブラウザを備えた情報処理装置と、当該情報処理装置におけるポータルサイトとを備えた情報抽出システムであって、

前記ポータルサイトは、前記情報処理装置からアクセスされた際に、前記情報処理装置においてローカルプロキシとして動作するプログラムファイルを前記情報処理装置に送信し、

前記情報処理装置のウェブブラウザは、前記ウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、

前記情報処理装置において動作する前記ローカルプロキシは、

前記ウェブブラウザの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された当該操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、前記ウェブコンテンツの中から当該特定の操作の対象となった情報を抽出する処理とを前記ウェブブラウザに実行させる、前記ウェブブラウザの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを、前記ウェブサーバから受信した前記ウェブコンテンツに埋め込むと共に、

前記ウェブブラウザにて抽出された前記情報を前記ポータルサイトに送信することを特徴とする情報抽出システム。

【請求項６】　ウェブコンテンツを格納するウェブサーバと、通信ネットワークを介して当該ウェブサーバからウェブコンテンツを受信して付加的な処理を行うプロキシサーバと、当該プロキシサーバにて当該付加的な処理を施された当該ウェブコンテンツを表示するクライアントとを備えた情報抽出システムであって、

前記クライアントは、前記ウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、

前記プロキシサーバは、

前記クライアントにて検出された前記操作イベントを収集する操作イベント取得手段と、

前記クライアントから受け取った前記操作イベントの列を解析して予め定められた特定の操作を抽出する操作イベント解析手段と、

前記ウェブコンテンツの中から当該特定の操作の対象となった情報を抽出する情報抽出手段とを備えることを特徴とする情報抽出システム。

【請求項７】　文書データを表示する閲覧手段と、

前記閲覧手段にて表示された文書を閲覧する際にユーザが行った入力操作に基づいて、ユーザが興味を持つ情報に対して無意識的に行った操作として定義付けた操作を検出する操作検出手段と、

前記操作検出手段により検出された、前記閲覧手段の表示画面上における前記特定の操作の行われた場所に表示されている文字列を抽出する文字列抽出手段とを備えることを特徴とする情報処理装置。

【請求項８】　前記文字列抽出手段は、前記特定の操作の行われた場所に表示されている文字列を含む文または行を単位として抽出を行うことを特徴とする請求項６に記載の情報処理装置。

【請求項９】　ウェブサーバからウェブコンテンツを受信して表示するウェブブラウザを備えた情報処理装置と接続し、当該情報処理装置の情報を収集する情報収集装置において、

前記情報処理装置において実行されることにより、前記ウェブブラウザの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された当該操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、前記ウェブコンテンツの中から当該特定の操作の対象となった情報を抽出する処理とを前記ウェブブラウザに実行させる、前記ウェブブラウザの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを、前記ウェブサーバから受信した前記ウェブコンテンツに埋め込むプログラムファイルを記憶した記憶手段と、

前記記憶手段から前記プログラムファイルを読み出して前記情報処理装置に送信する送信手段と、

前記情報処理装置にて抽出された前記情報を収集する情報収集手段とを備えることを特徴とする情報収集装置。

【請求項１０】　前記情報収集装置の前記記憶手段に記憶された前記プログラムファイルは、Ｊａｖａアプレットにて作成され、前記プログラムパッケージ

をＪａｖａスクリプトで記述して前記ウェブコンテンツに埋め込むことを特徴とする請求項８に記載の情報収集装置。

【請求項１１】　文書データを表示する表示画面上でユーザが行った入力操作に基づいて予め定められた特定の操作を検出するステップと、

検出された、前記表示画面上における前記特定の操作の行われた場所に表示されている文字列を、当該文字列を含む文または行を単位として抽出するステップとを含むことを特徴とする文字列抽出方法。

【請求項１２】　文書データを表示する表示画面上でユーザが行った入力操作に基づいて、ポインティングデバイスのポインタを、表示されている文書の行に沿って移動させるなぞり読み動作を検出するステップと、

検出された、前記表示画面上における前記なぞり読み動作の行われた場所に表示されている文字列を、当該文字列を含む文または行を単位として抽出するステップとを含むことを特徴とする文字列抽出方法。

【請求項１３】　前記文字列を抽出するステップにおいて、前記文書の表示画面における前記ポインタが移動した位置の文字列の１行上の文字列を含む文または行を抽出することを特徴とする請求項１１に記載の文字列抽出方法。

【請求項１４】　文書データを表示する表示画面上でユーザが行った入力操作に基づいて、ポインティングデバイスのポインタを、表示されている文書の行を順に指しながら当該行とは直交する方向に移動させる行単位なぞり読み動作を検出するステップと、

検出された、前記表示画面上における前記行単位なぞり読み動作の行われた場所に表示されている文字列を、当該文字列を含む文または行を単位として抽出するステップとを含むことを特徴とする文字列抽出方法。

【請求項１５】　コンピュータに実行させるプログラムを当該コンピュータの入力手段が読取可能に記憶した記憶媒体において、

前記プログラムは、

文書データを表示する処理と、

文書データを表示する表示画面上でユーザが行った入力操作に基づいて予め定められた特定の操作を検出する処理と、

検出された、前記表示画面上における前記特定の操作の行われた場所に表示されている文字列を抽出する処理とを前記コンピュータに実行させることを特徴とする記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、コンピュータのユーザによるディスプレイ装置の画面上での操作を監視して情報を取得する情報処理方法に関する。

【０００２】

【従来の技術】

近年、オンラインショッピング、バナー広告等のようなインターネットの商用的な利用が盛んになりつつある。そこで、この種のインターネットの利用方法の効果を上げるため、ウェブサイトの運営者は、ウェブコンテンツに対するユーザの反応（ウェブ視聴率）の調査を行い、ウェブコンテンツやウェブサイトのデザインに反映させたり、One-to-Oneマーケティングに活用するといったことを行っている。

既存のウェブ視聴率調査の手法では、ウェブサイトの中でユーザがどのような話題やテーマに注目したかという情報を獲得しようとする場合、ユーザからアンケートを取る方法や、ページの表示時間及びページの訪問回数等のアクセス情報を取得して、得られた情報を基に推定する方法があった。ここで用いられるアクセス情報としては、サーバ側で得られるＨＴＴＰ要求の数（ヒット数）や、クライアント側で得られる特定のウェブコンテンツの閲覧に関する情報がある。

【０００３】

ウェブ視聴率調査を行うためにアンケートを取る方法では、ユーザに明示的な入力をしてもらうことによって調査を実施する。具体的には、例えば、ウェブコンテンツに予めアンケートのページを用意して、興味のあるトピックやキーワードについて答えてもらったり、ウェブページに「面白かった」「つまらなかった」などの感想を示すボタンを設けておき、ユーザにクリックしてもらうといった方法が採られる。

この方法では、ユーザによる明示的な入力操作の結果として情報を得られるため、ユーザの抱く興味の傾向などに関して、得られた情報に高い信頼性を期待できる。

【０００４】

ウェブ視聴率調査で用いるためにサーバ側で取得できる情報としては、ウェブコンテンツへのＨＴＴＰ要求の数（ヒット数）がある。ウェブブラウザでウェブページの１ページを読む場合、当該ウェブページ中に画像などが埋め込まれたり、フレームが使われている場合は、当該１ページに対するヒット数は複数となる。また、ウェブサーバは、当該ウェブコンテンツから他のウェブコンテンツへ遷移したときのＨＴＴＰ要求は受け取らない。

この方法では、ユーザがどの資源（ウェブコンテンツ）にアクセスしたかについては全て記録できる。そして、資源の種別（ＨＴＭＬ、画像などの別）等の情報と組み合わせることにより、所定のウェブコンテンツへの滞在時間を推測することができる。

【０００５】

クライアント側では、ウェブブラウザにて表示されたウィンドウの状態を監視することができるため、サーバに比べて更に詳細な情報を取得することができる。例えば、ページ単位で表示時間を測定したり、ウィンドウの位置、大きさやフォーカスの変化を調べたり、ユーザによるキー入力からキーワードを記録したりすることができる。また、特定のウェブサイトに縛られることなくユーザの閲覧履歴を記録することができる。

これらの手法で得られた情報に基づいて、ユーザの抱く興味の傾向などに関して、ある程度推測することができる。

【０００６】

また、ユーザが必要とする情報の獲得を支援するサービスとして検索エンジンがある。検索エンジンでは、ユーザがキーワードを入力することにより、当該キーワードを含むページが検索され、提示される。しかし、検索エンジンで索引可能なページ数は膨大であるため、適切に検索数を絞り込むことが重要である。検索数の絞り込みの技術として、検索結果に対してユーザに自分の興味に適合する

ページを指摘させ、指摘されたページの中のキーワードを用いて再度検索を行う手法が取られていた。この場合、検索結果に対してユーザが検索目的に適合しているとして選択した文書中のキーワードを利用して、検索条件の自動変更を行う。これにより、ユーザの抱く興味の傾向などを検索結果に反映させることができる。なお、ここでは、指摘されたページ全体に含まれるキーワードを利用する。

【０００７】

【発明が解決しようとする課題】

しかし、上記のような従来のウェブ視聴率調査や検索エンジンにおいて、ユーザがどのような話題やテーマに注目したかといったユーザの興味に関する情報を取得する場合、得られる情報量や信頼性の点で十分とはいえなかった。

ユーザに対してアンケートに答えてもらう方法では、ユーザにとってはアンケートに答えるという負担があるため、高い回答率が得られない場合があった。同様に、ユーザの負担を考慮すると、ページ中の文単位のような細かい項目ごとの評価を要求するといった煩雑なアンケートを行うことは困難である。さらに、アンケートを行うためには、アンケート用のページやボタンを用意するなどの措置が必要であるため、任意のウェブコンテンツに関する情報を容易に取得することができなかった。

【０００８】

サーバマシンやクライアントマシンにおいて得られる情報を用いて推測する方法では、上記のように、サーバにおいて得られる情報は所定のウェブコンテンツへのヒット数である。そのため、所定のウェブコンテンツへの滞在時間を推測することはできるが、ユーザがどのウェブページをどの程度の時間読んだかというようなウェブページごとの詳細な情報を得ることはできない。

このような情報は、クライアントマシンにおいてウェブブラウザにて表示されたウィンドウの状態を監視することにより得ることができる。しかし、ウェブブラウザを監視する手段は、アプリケーションプログラムやプロキシサーバとしてクライアントマシンに実装され、ウェブブラウザの外で動作するため、ウェブページのデータ構造にアクセスすることができない。そのため、マウス操作の操作対象となるＨＴＭＬ中のオブジェクトを記録することはできない。したがって、

ユーザがウェブコンテンツの中のどの部分に特に興味を持ったかというような更に詳細な情報を得ることはできない。

【０００９】

ウェブブラウザ中の情報にアクセスする方法としては、ウェブコンテンツ自体をＪａｖａアプレットによって実装する方法が考えられる。しかし、この方法であっても、当該Ｊａｖａアプレット中の内容しかわからないため、一般のウェブページには適用できない。

【００１０】

また、検索エンジンにおいてユーザの評価に基づいて検索条件を変更する方法は、上述したように、検索条件を変更するために用いるキーワードが対象であるウェブページの文書全体から抽出されたものである。そのため、ウェブページの文書中においてユーザが特に注目した部分（文やキーワード）等を、検索条件にきめ細かく反映させることができない。

【００１１】

本発明は以上のような技術的課題を解決するためになされたものであって、ユーザによる明示的な入力を必要とせず、かつウェブコンテンツにおいてユーザが興味を持った箇所に関する詳細な情報を取得できるようにする。

また、ユーザによるウェブブラウザ上での操作を含む詳細な操作を抽出して、ユーザの抱く興味の傾向を示す情報として利用できるようにする。

【００１２】

【課題を解決するための手段】

かかる目的のもと、本発明は、通信ネットワークにて接続されたサーバとクライアントとを備えた情報抽出システムであって、サーバは、クライアントにおいて閲覧に供されるデータファイルを提供し、クライアントは、通信ネットワークを介してサーバから受信したこのデータファイルの内容を表示する閲覧手段と、この閲覧手段にて表示されたデータファイルの内容を閲覧する際にユーザが行った入力操作に基づいて予め定められた特定の操作を検出する操作検出手段と、この操作検出手段により検出された、閲覧手段の表示画面上における特定の操作の行われた場所に表示されている情報を抽出する手段とを備えることを特徴として

いる。

【００１３】

また本発明は、ウェブコンテンツを格納するウェブサーバと、通信ネットワークを介して当該ウェブサーバからウェブコンテンツを受信して表示するクライアントとを備えた情報抽出システムであって、クライアントは、このウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、ウェブサーバに格納されたこのウェブコンテンツは、このクライアントの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、このウェブコンテンツの中からこの特定の操作の対象となった情報を抽出して前記ウェブサーバに返送する処理とをこのクライアントに実行させる、このクライアントの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを埋め込んであることを特徴としている。このような構成とすれば、ウェブコンテンツの作成者がウェブコンテンツにこのようなプログラムパッケージを埋め込んでおくことにより、このウェブコンテンツが情報処理装置からアクセスされることによって、かかる情報処理装置においてユーザが興味を示した内容に関する情報を取得できる点で優れている。得られた情報は、ウェブ視聴率調査や検索エンジンにおける検索条件の絞り込みといったサービスに利用することができる。

【００１４】

これとは異なり、ウェブコンテンツを格納するウェブサーバと、通信ネットワークを介して当該ウェブサーバからウェブコンテンツを受信して表示するクライアントとを備えた情報抽出システムであって、クライアントは、このウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、ウェブサーバは、このクライアントの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、このウェブコンテンツの中からこの特定の操作の対象となった情報を抽出して前記ウェブサーバに返送する処理とをこのクライアントに実行させる、このクライアントの機能を拡張するための機能拡張

プログラム言語にて記述されたプログラムパッケージをこのウェブコンテンツに埋め込んでクライアントに送信することを特徴としている。このような構成とすれば、ウェブサーバにおいて、格納されているウェブコンテンツに対してユーザが興味を示した情報を取得できる点で優れている。得られた情報は、ウェブ視聴率調査や検索エンジンにおける検索条件の絞り込みといったサービスに利用することができる。

【００１５】

さらに、ウェブコンテンツを格納するウェブサーバと、通信ネットワークを介してこのウェブサーバからウェブコンテンツを受信して付加的な処理を行うプロキシサーバと、このプロキシサーバにてこの付加的な処理を施されたこのウェブコンテンツを表示するクライアントとを備えた情報抽出システムであって、クライアントは、このウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、プロキシサーバは、クライアントの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、このウェブコンテンツの中からこの特定の操作の対象となった情報を抽出して前記プロキシサーバに返送する処理とをこのクライアントに実行させる、このクライアントの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを、ウェブサーバから受信したウェブコンテンツに埋め込んでクライアントに送信することを特徴としている。このような構成とすれば、プロキシサーバにおいて、ウェブコンテンツを受信した情報処理装置のユーザがこのウェブコンテンツにおける興味を示した内容に関する情報を取得できる点で優れている。得られた情報は、ウェブ視聴率調査や検索エンジンにおける検索条件の絞り込みといったサービスに利用することができる。

【００１６】

ここで、プロキシサーバは、前記のようなプログラムパッケージをクライアントに送るのではなく、クライアントにて検出された操作イベントを収集する操作イベント取得手段と、このクライアントから受け取ったこの操作イベントの列を解析して予め定められた特定の操作を抽出する操作イベント解析手段と、このウ

ェブコンテンツの中からこの特定の操作の対象となった情報を抽出する情報抽出手段とを備えることを特徴とすることができる。このように構成すれば、操作イベントに基づく特定の操作の抽出及びユーザが興味を持った情報の抽出をプロキシサーバにおいて実行するため、クライアントにおける負担を軽減させることができる点で好ましい。プロキシサーバが情報を抽出する対象であるウェブコンテンツは、プロキシサーバがウェブサーバから受信したものを保持しておいて用いても良いし、情報抽出の際にクライアントに要求しても良い。

【００１７】

さらにまた、ウェブコンテンツを格納するウェブサイトと、通信ネットワークを介してこのウェブサイトからウェブコンテンツを受信して表示するウェブブラウザを備えた情報処理装置と、この情報処理装置におけるポータルサイトとを備えた情報抽出システムであって、このポータルサイトは、情報処理装置からアクセスされた際に、この情報処理装置においてローカルプロキシとして動作するプログラムファイルをこの情報処理装置に送信し、この情報処理装置のウェブブラウザは、ウェブコンテンツの表示画面においてユーザが行った操作を操作イベントとして検出する機能を備え、情報処理装置において動作するこのローカルプロキシは、ウェブブラウザの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、このウェブコンテンツの中からこの特定の操作の対象となった情報を抽出する処理とをこのウェブブラウザに実行させる、このウェブブラウザの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを、ウェブサーバから受信したこのウェブコンテンツに埋め込むと共に、このウェブブラウザにて抽出された情報を前記ポータルサイトに送信することを特徴としている。このような構成とすれば、ポータルサイトにおいて、情報処理装置のユーザが受信したウェブコンテンツにおける興味を示した内容に関する情報を取得できる点で優れている。得られた情報は、ウェブ視聴率調査や検索エンジンにおける検索条件の絞り込みといったサービスに利用することができる。

【００１８】

また、本発明は、文書データを表示する閲覧手段と、この閲覧手段にて表示さ

れた文書を閲覧する際にユーザが行った入力操作に基づいて、ユーザが興味を持つ情報に対して無意識的に行った操作として定義付けた操作を検出する操作検出手段と、この操作検出手段により検出された、この閲覧手段の表示画面上におけるこの特定の操作の行われた場所に表示されている文字列を抽出する文字列抽出手段とを備えることを特徴としている。ユーザが興味を持つ情報に対して無意識的に行った操作とは、アンケートへの入力のような積極的な操作とは異なり、ユーザが文書の内容を注意しながら読む際に、マウスポインタでテキストをなぞりながら読んだり、読んでいる範囲を範囲選択して読んだりするような無意識的に現れる動作である。これらの動作を検出してその動作の対象となった情報を取得することにより、ユーザの積極的な入力操作を要請することなくユーザが興味を示した内容に関する情報を取得できる点で好ましい。

【００１９】

ここで文字列抽出手段は、この特定の操作の行われた場所に表示されている文字列を含む文または行を単位として抽出を行うことを特徴としている。文または行を単位として抽出するには、抽出する文字列の範囲を拡張して、改行コードや「。」等の文または行の区切りを示す箇所を検出し、その範囲でテキストを抽出する。これにより、ユーザが興味を示した内容を、まとまった意味のある情報として抽出することができる点で好ましい。

【００２０】

また、本発明はウェブサーバからウェブコンテンツを受信して表示するウェブブラウザを備えた情報処理装置と接続し、この情報処理装置の情報を収集する情報収集装置において、この情報処理装置において実行されることにより、このウェブブラウザの操作イベント検出機能を用いて操作イベントを検出する処理と、検出された操作イベントの列を解析して予め定められた特定の操作を抽出する処理と、このウェブコンテンツの中からこの特定の操作の対象となった情報を抽出する処理とをこのウェブブラウザに実行させる、このウェブブラウザの機能を拡張するための機能拡張プログラム言語にて記述されたプログラムパッケージを、ウェブサーバから受信したウェブコンテンツに埋め込むプログラムファイルを記憶する記憶手段と、このプログラムファイルを記憶手段から読み出して情報処理

装置に送信する送信手段と、情報処理装置にて抽出されたこの情報を収集する情報収集手段とを備えることを特徴としている。

【００２１】

ここで、前記情報収集装置の前記記憶手段に記憶された前記プログラムファイルは、Ｊａｖａアプレットにて作成され、前記プログラムパッケージをＪａｖａスクリプトで記述して前記ウェブコンテンツに埋め込むことを特徴としている。このように構成すれば、パーソナルコンピュータなどで広く利用されているＪａｖａに対応したウェブブラウザを用いて情報の抽出を行うことができる点で好ましい。また、Ｊａｖａアプレットにて作成することにより、予めプログラムファイルを情報処理装置に配布しておく必要がない点で優れている。

【００２２】

また、本発明は、文書データを表示する表示画面上でユーザが行った入力操作に基づいて予め定められた特定の操作を検出するステップと、検出された、この表示画面上におけるこの特定の操作の行われた場所に表示されている文字列を、この文字列を含む文または行を単位として抽出するステップとを含むことを特徴としている。

【００２３】

更に、本発明は、文書データを表示する表示画面上でユーザが行った入力操作に基づいて、ポインティングデバイスのポインタを、表示されている文書の行に沿って移動させるなぞり読み動作を検出するステップと、検出された、この表示画面上におけるこのなぞり読み動作の行われた場所に表示されている文字列を、この文字列を含む文または行を単位として抽出するステップとを含むことを特徴としている。このような構成とすれば、なぞり読みが行われた時点で、ユーザによる積極的な入力操作を必要とせずに、このなぞり読みの行われた箇所のテキストを抽出できる点で特に優れている。また、文または行単位で文字列を抽出することにより、ユーザが興味を示した内容を、まとまった意味のある情報として抽出することができる点で好ましい。

【００２４】

ここで、この文字列を抽出するステップにおいて、文書の表示画面におけるポ

インタが移動した位置の文字列の１行上の文字列を含む文または行を抽出することを特徴とすることができる。ユーザはなぞり読みを行っている際に、マウスポインタのある行の一つ上の行を読む場合がある。そこで、このような構成とすれば、ユーザが興味を持って読んだと推定される情報を漏らさず抽出することができる点で優れている。

更にまた、本発明は、文書データを表示する表示画面上でユーザが行った入力操作に基づいて、ポインティングデバイスのポインタを、表示されている文書の行を順に指しながら当該行とは直交する方向に移動させる行単位なぞり読み動作を検出するステップと、検出された、この表示画面上におけるこの行単位なぞり読み動作の行われた場所に表示されている文字列を、この文字列を含む文または行を単位として抽出するステップとを含むことを特徴としている。このような構成とすれば、ユーザがマウスをテキストの行に直交する方向に動かしながら、長い文章を読んだ場合に、ユーザによる積極的な入力操作を必要とせずに、この行単位なぞり読みの行われた箇所のテキストを抽出できる点で特に優れている。

なお、横書きのテキストにおいては、なぞり読みは行と一致する横方向のポインタの動きから検出し、行単位なぞり読みは行と直交する縦方向のポインタの動きから検出する。これに対し、縦書きのテキストにおいては、なぞり読みは行と一致する縦方向のポインタの動きから検出し、行単位なぞり読みは行と直交する横方向のポインタの動きから検出する。

また、本発明は、コンピュータに実行させるプログラムを当該コンピュータの入力手段が読取可能に記憶した記憶媒体において、このプログラムは、文書データを表示する処理と、この文書データを表示する表示画面上でユーザが行った入力操作に基づいて予め定められた特定の操作を検出する処理と、検出された、この表示画面上におけるこの特定の操作の行われた場所に表示されている文字列を抽出する処理とをこのコンピュータに実行させることを特徴としている。このように構成すれば、このプログラムをロードした情報処理装置において文書データを表示した場合に、ユーザがこの文書における興味を示した内容に関する情報を取得できる点で優れている。得られた情報は、ウェブ視聴率調査や検索エンジンにおける検索条件の絞り込みといったサービスを提供するサーバに転送すれば、

これらのサービスに利用することができる。

【００２５】

【発明の実施の形態】

以下、添付図面に示す実施の形態に基づいてこの発明を詳細に説明する。

まず、本発明の概要を説明する。本発明は、ユーザがコンピュータの画面上で文書を閲覧する際の無意識的なマウスの動きと、ユーザの興味の対象との間に関連があると考え、マウスの特徴的な動作を検出することにより、ユーザが興味を示したと推定される情報を抽出する。マウスの動作に基づいてユーザの興味に関わる情報を抽出するため、文書中の文や単語、あるいは挿入されている図表といった小さな単位でユーザが注目した対象を特定することができる。

【００２６】

本実施の形態では、以下にあげる五つのマウスの動きを、ユーザが興味を持って注目した対象に対して無意識的に行った操作と定義する。

１．マウスのボタンを押したままマウスポインタを移動させる（ドラッグする）ことにより、マウスポインタで指定された範囲のテキストを選択する動作（以下、テキスト選択と称す）。

２．マウスポインタをリンクの上に重ねるリンクへのポインティング動作。

３．マウスによるリンクのクリック。

４．マウスポインタをテキストの行に沿って横方向に動かしながらその行を読むときの、当該マウスポインタを横方向に移動させる動作（以下、なぞり読みと称す）。

５．テキスト中の読んでいる行をマウスポインタで指し、読む行を移動するのに伴って、マウスポインタを縦に少しずつ動かす動作（以下、縦方向なぞり読みと称す）。

なお、ここではポインティングデバイスとしてマウスを用いた場合の動作を定義したが、トラックボールやペンタブレットなどの他のポインティングデバイスを用いた場合でも、ユーザが興味を持った対象に対しては、概ね同様の動作を行うと考えられる。したがって、以下の説明においても、特にポインティングデバイスの種類を区別することなく、マウスを用いる場合を例として説明する。

また、ユーザが興味を持って注目した対象に対して無意識的に行った操作は上の五つの操作に限定されるものではない。この他にも、経験的に、ユーザが興味を持って注目した対象に対して行ったと推定される操作を任意に定義することによって、情報の抽出に用いることができる。

【００２７】

図１は、本実施の形態における情報抽出システムの全体構成を説明する図である。同図において、符号１０は操作イベント検出部であり、コンピュータの画面に表示された文書上におけるマウスの動作を監視して、その中から操作イベントを検出する。符号２０は操作イベント解析部であり、操作イベント検出部１０により検出された操作イベントの列（以下、操作イベント列と称す）を解析して、ユーザが興味を示した対象に対して行ったと推定される特定の操作を抽出する。符号３０はテキスト抽出部であり、コンピュータの画面に表示された文書のうちで、操作イベント解析部２０により抽出された操作の対象となったテキストを抽出する。これらの各構成要素は、コンピュータに上記各処理を実行させるためのプログラムモジュールとして実現される。

なお、本実施の形態では、インターネットなどで一般的に用いられているウェブコンテンツを表示するためのウェブブラウザの表示画面を、マウスの動作を監視する対象領域とする。すなわち、ウェブブラウザにて表示されたウェブコンテンツ（ホームページ等）上におけるマウスの動作から操作イベント検出部１０が操作イベントを検出し、検出した操作イベントから操作イベント解析部２０が特定の操作を抽出し、テキスト抽出部３０がかかる操作の対象となったテキストをユーザが興味を持って注目した情報として抽出する。この場合、操作イベント検出部１０，操作イベント解析部２０及びテキスト抽出部３０は、ダイナミックＨＴＭＬの機能を用いて実現することができる。

【００２８】

操作イベント検出部１０は、JavaScript等のスクリプト言語を用いてＨＴＭＬファイルに埋め込む形態で実現することができる。JavaScriptでは、マウスの移動やクリック、ドラッグ、文字列の選択、キーを押下したり離したりする動作、画面のスクロール等の操作を、イベントとして抽出することが可能である。例え

ば、マウスの移動に対してonMouseMoveというイベントハンドラを定義し、ＨＴＭＬファイルに記述しておくことにより、マウスが移動した場合に、その動作を操作イベントとして検出することができる。なお、ウェブコンテンツ以外の、所定のアプリケーションプログラムにて作成された文書を対象として、当該文書を表示している表示画面上におけるマウスの動作を監視する場合でも、オペレーティングシステムのＡＰＩ等を利用することにより、マウスの特定のから操作イベントを抽出することができる。

【００２９】

操作イベント解析部２０は、操作イベント検出部１０により検出された操作イベント列を解析し、当該操作イベント列が予め定義されている特定の操作に該当するかどうかを調べる。そして、所定の操作イベント列が当該特定の操作に該当する場合、当該操作が行われたことをテキスト抽出部３０に通知する。また、テキスト抽出部３０におけるテキストの抽出に用いるため、当該操作が行われた位置（座標値）等の具体的な情報をテキスト抽出部３０に送る。ここで、予め定義された特定の操作とは、ユーザが興味を持った対象に対して無意識的に行うと推定された操作である。本実施の形態では、上述した五つの操作、すなわち、１．テキスト選択、２．リンクへのポインティング、３．リンクのクリック、４．なぞり読み、５．縦方向なぞり読みを特定の操作として定義する。操作イベント検出部１０により検出された操作イベント列の中からこれらの特定の操作を抽出する処理の詳細については後述する。

【００３０】

テキスト抽出部３０は、操作イベント解析部２０から特定の操作を抽出したことを示す通知を受け取ると、操作イベント解析部２０からさらにテキストの抽出に必要な座標値などの情報を受信する。そして、受け取った情報を用いて、ウェブブラウザにて表示されているウェブコンテンツの該当個所から、当該特定の操作の対象となったテキストを取得する。操作イベント解析部２０によって抽出される特定の操作ごとにおけるテキストの抽出処理の詳細については後述する。

【００３１】

なお、取得したテキストは、当該テキストを利用する他のシステムに送信され

る。例えば、ウェブ視聴率調査を行うシステムや検索エンジンにおいて、テキスト抽出部３０により取得されたテキストを受信し、ユーザが興味を持った対象に関する情報として利用することができる。

【００３２】

図２は、操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０の働きを概念的に説明する図である。図２において、操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０は、JavaScriptで記述され、ウェブコンテンツ２００に埋め込まれているものとする。

図２を参照すると、まず、ウェブブラウザにて表示されたウェブコンテンツ２００のうち、所定のテキスト２０１に対してマウスによる特定の操作が行われたものとする（２１１）。すると、操作イベント検出部１０により、マウスの動作に基づく操作イベントが検出される。検出された操作イベントは操作イベント解析部２０に送られる（２１２）。次に、操作イベント解析部２０により、操作イベント列の解析が行われて、特定の操作が抽出される。そして、特定の操作が抽出されたことを示す通知及び操作の内容に関する情報がテキスト抽出部３０に送られる（２１３）。この後、テキスト抽出部３０により、抽出された特定の操作に応じた処理が行われて、ウェブコンテンツ２００の中からテキスト２０１が抽出される（２１４）。

【００３３】

こうして得られたテキスト２０１は、ユーザがウェブコンテンツ２００を閲覧する際に、興味を持った情報であると考えられるため、ウェブ視聴率の調査や、検索エンジンにおける検索条件の絞り込み等、種々のサービスに利用することができる。そこで、抽出されたテキスト２０１は、当該テキスト２０１をユーザに関する情報として利用しようとする利用者に送られる必要がある。テキスト２０１の送信は、スクリプトの形でウェブコンテンツ２００に埋め込んでおき、ウェブブラウザの機能を用いて行ったり、所定のプログラムを情報処理装置に提供しておき、当該プログラムの機能にて行う等、種々の方法を採ることができる。

【００３４】

次に、個々の特定の操作ごとに、本実施の形態によるテキストを取得する処理

を詳細に説明する。

まず、特定の操作としてテキスト選択が行われた場合について説明する。

操作イベント解析部２０は、操作イベント検出部１０から送られてくる操作イベント列の中から、テキストを選択する操作を行った時に発生するselectイベントを検出すると、このselectイベントに基づいて、テキスト選択の操作に対応するselectionオブジェクトを取得する。テキスト選択の捜査の終了は、押下されているマウスボタンを離した時に発生するmouseupイベントによって認識することができる。

なお、ダイナミックＨＴＭＬでは、テキスト選択の操作を行うと、選択された範囲をselectionオブジェクトとして取得することができる。したがって、ダイナミックＨＴＭＬに対応したウェブブラウザにおいては、ユーザによるテキスト選択の操作に応じて直ちにselectionオブジェクトが取得されることとなる。

【００３５】

テキスト抽出部３０は、操作イベント解析部２０にて生成されたselectionオブジェクトを用いて、選択されたテキストを抽出する。これにより、図４に示すように、「This cat is very smart.」という文の中から「cat is very」という文字列が抽出される。そして、抽出された文字列「cat is very」が所定のシステムに送られ、ユーザが興味を持って注目した情報として利用される。

【００３６】

図３は、上述したテキスト抽出部３０によるテキスト抽出処理を、ダイナミックＨＴＭＬを用いて実現する場合のプログラムを説明する図である。図４は、当該プログラムにより、「This cat is very smart.」という文の中から「cat is very」というテキスト列が選択された場合のテキスト抽出の課程を説明する図である。

ここでは、テキストを抽出するための関数としてgetSelectedText関数を定義する。getSelectedText関数の引数は、ユーザのテキスト選択により発生したselectionオブジェクトslである（図４のselectionオブジェクト４０１）。

図３のプログラムリストの３行目において、取得したselectionオブジェクトslに基づいて、createRangeメソッドによりTextRangeオブジェクトtrを生成する

（図４のTextRangeオブジェクト４０２）。ここで、TextRangeオブジェクトとは、ダイナミックＨＴＭＬでテキスト操作を行うためのオブジェクトである。

次に、プログラムリストの４行目において、TextRangeオブジェクトのtextプロパティにより、選択範囲のテキスト「cat is very」を抽出する（図４のテキスト４０３）。

【００３７】

次に、特定の操作としてリンクのポインティングが行われた場合について説明する。

操作イベント解析部２０は、操作イベント検出部１０から送られてくる操作イベント列の中から、マウスポインタがリンクの上に乗った時に発生するイベントと、リンクから外れた時に発生するイベントを用いてリンクのポインティング操作を検出する。本実施の形態では、これらのイベント発生と同時に、リンクタグの付けられているテキストとリンクが張られた部分を持つテキストとを文または行の単位で抽出する。また、マウスポインタが単にリンクの上を通過した場合と、偶然リンクの上でマウスポインタが長時間停止した場合とを排除するため、ポインティングしていた時間を測定して判断材料とする。

【００３８】

具体的には、まず、マウスポインタがリンクの上に乗ったことを示すイベント（mouseoverイベント）が発生すると、その発生時刻ｔ１を記憶する。その後、マウスの移動イベント（mousemoveイベント）が発生すると、リンク上のマウスポインタの位置（座標値）を求める。次に、マウスポインタがリンクから外れたことを示すイベント（mouseoutイベント）が発生したならば、その発生時刻ｔ２を求める。そして、しきい値T_l、T_hに対して、$T_l < (t2 - t1) < T_h$であれば、マウスによるリンクのポインティング操作が行われたものと認識してテキスト抽出部３０に通知し、mousemoveイベントにより得られたマウスポインタの位置情報をテキスト抽出部３０に送る。

ここで、しきい値T_l、T_hはユーザが無意識にマウスポインタをリンク上に通過させた場合と偶然マウスポインタがリンク上で長時間停止した場合とを排除するために設けられたものである。すなわち、$T_l \geqq (t2 - t1)$の場合は、マ

ウスポインタがリンク上を通過したに過ぎないと判断して、テキスト抽出部３０への通知は行わない。また、（ｔ２－ｔ１）≧T_hの場合は、偶然マウスポインタがリンク上に長時間位置したものと判断して、テキスト抽出部３０への通知は行わない。

【００３９】

テキスト抽出部３０は、リンクのポインティング操作が行われたことを知らせる通知を受け取ると、当該ポインティング操作が行われた箇所がリンクタグであれば、当該リンクタグの付けられたテキストを文または行の単位で抽出する。また、当該ポインティング操作が行われた箇所が文中の所定の箇所に張られたリンクであれば、当該リンクを含むテキストを文または行の単位で抽出する。

【００４０】

ここで、文または行の単位でテキストを抽出する方法について説明する。テキストを文または行の単位で区切るには、まず、ポインティング操作の対象となったリンクタグまたはポインティング操作が行われた位置（座標値）から抽出対象のテキストの範囲を順次広げていく。そして、改行コード、「。」「、」「．」等の文や行の区切りを示す記号が現れたところでテキストの範囲の拡張を止め、得られたテキスト列を抽出する。

【００４１】

図５は、上述したテキスト抽出部３０によるテキスト抽出処理を、ダイナミックＨＴＭＬを用いて実現する場合のプログラムを説明する図である。図６は、当該プログラムにより、ウェブブラウザにて表示されたウェブコンテンツの文書の中から「This cat is very smart」という文のなかのリンク（下線が引かれた「cat」の部分）にポインティングが行われた場合のテキスト抽出の課程を説明する図である。

ここでは、テキストを抽出するための関数としてgetLinkTagText関数とgetLinkText関数とを定義する。

【００４２】

getLinkTagText関数は、リンクタグの付けられたテキストを抽出する関数である。引数はアンカーオブジェクトanchorである。そして、図５のプログラムリス

トの３行目において、当該リンクタグの付けられているテキスト全部の抽出を行っている。

getLinkText関数は、リンクを張られた部分を持つテキストを文または行の単位で抽出する関数である。引数はマウスポインタの存在する座標である。以下、図６を参照しながら、getLinkText関数によるテキストの抽出の処理を説明する。

図５のプログラムリストの８行目において、bodyオブジェクトに対してcreateTestRangeメソッドを用い、当該ウェブコンテンツのページ全体を含むTextRangeオブジェクトを生成する（図６のTextRangeオブジェクト６０１）。

次に、プログラムリストの９行目において、moveToPointメソッドにより、TextRangeオブジェクトにマウスポインタが指している文字を指し示すようにする（図６のTextRangeオブジェクト６０２）。

次に、プログラムリストの１０行目において、テキストの選択領域を変更する関数（図６のchangeTextRange関数は、この処理を行うために定義した関数である）により、TextRangeオブジェクトの選択範囲を文または行の単位まで拡張する（図６のTextRangeオブジェクト６０３）。

最後に、プログラムリストの１１行目において、TextRangeオブジェクトのTextプロパティにより、テキスト「This cat is very smart.」を抽出する（図６のテキスト６０４）。

【００４３】

次に、特定の操作としてリンクのクリックが行われた場合について説明する。

操作イベント解析部２０は、操作イベント検出部１０から送られてくる操作イベント列の中から、リンクをクリックしたときに発生するイベントを用いてリンクのクリック操作を検出する。本実施の形態では、リンクのポインティングの場合と同様に、イベント発生と同時に、リンクタグの付けられているテキストとリンクが張られた部分を持つテキストとを文または行の単位で抽出する。

具体的には、マウスポインタがリンクの上に乗ったことを示すイベント（mouseoverイベント）が発生すると、その発生時刻ｔ１を記憶する。その後、マウスの移動イベント（mousemoveイベント）が発生すると、リンク上のマウスポイン

タの位置（座標値）を求める。次に、クリックイベント（clickイベント）が発生したならば、テキスト抽出部３０に通知し、mousemoveイベントにより得られたマウスポインタの位置情報をテキスト抽出部３０に送る。

【００４４】

テキスト抽出部３０は、リンクのクリック操作が行われたことを知らせる通知を受け取ると、当該クリック操作が行われた箇所がリンクタグであれば、当該リンクタグの付けられたテキストを文または行の単位で抽出する。また、当該クリック操作が行われた箇所が文中の所定の箇所に張られたリンクであれば、当該リンクを含むテキストを文または行の単位で抽出する。

テキスト抽出部３０によるテキスト抽出の処理は、上述したポインティング操作の場合と同様であるため、説明を省略する。

【００４５】

次に、特定の操作としてなぞり読みが行われた場合について説明する。

なぞり読みの抽出は、マウスの移動イベントで得られるマウスポインタの位置（座標値）とイベント発生時刻とを利用して行う。なぞり読みの動きはマウスの水平方向の直線上の動きであり、これを検出する方法はいくつか考えられるが、本実施の形態では以下のような手法を用いる。

まず、マウスの動きに途切れのない水平方向の連続移動を検出する。そして、水平方向への連続移動の長さが所定のしきい値以上の場合になぞり読みとして検出する。これは、偶然マウスが水平方向の直線移動をした動作を排除するためである。このような場合に発生するマウスの水平方向の移動は、あまり長くならないと考えられるので、適当なしきい値を設定することにより排除することができる。なお、水平方向への連続移動は、マウスの移動イベントが発生するたびに、次の条件を判断することにより検出できる。

第１に、最も新しい数回分（２～４回程度）のマウスポインタの座標から得られる移動の傾きからマウスが表示画面上で水平方向の移動をしているかどうかを判断する。

第２に、最後のイベント発生時刻とその直前のイベント発生時刻との差からマウスの動きが途切れていないかどうかを判断する。

以上の条件を満足して、マウスが水平方向の移動をし、かつ動きが途切れていないと判断された場合、マウスは水平方向への連続移動をしていると認識する。そして、この二つの条件のどちらかが満たされなくなった場合、水平方向への連続移動が終了したと判断する。

【００４６】

以上の前提に基づいて、操作イベント解析部２０がなぞり読みの操作を検出する処理について説明する。以下の説明において、パラメータＡｒは、マウスの移動方向を水平方向とみなすかどうかを決定するための傾きに関するしきい値である。パラメータＴｒは、マウスの動作が連続移動であるかどうかを決定するための停止時間に関するしきい値である。また、パラメータＬは、検出された水平方向への連続移動をなぞり読みと判断するための移動の長さに関するしきい値である。なお、座標は直交するｘ－ｙ座標にて表され、ｘ方向が表示画面の水平方向（すなわち、行と平行な方向）、ｙ方向が表示画面の垂直方向（すなわち、行と直交する方向）とする。

操作イベント解析部２０は、mousemoveイベントが発生するたびに、当該マウスポインタの座標（x_i，y_i）とｎ回前に発生したmousemoveイベントにおけるマウスポインタの座標（x_{i-n}，y_{i-n}）との差（$x_i - x_{i-n}$，$y_i - y_{i-n}$）を求める。そして、ｘ方向（水平方向）の差が正の値であれば、傾きａを下式

$$a = (y_i - y_{i-n}) / (x_i - x_{i-n})$$

で求める。また、最後のイベント発生時刻t_iとその直前のイベント発生時刻t_{i-1}との間の時間間隔t_dを下式

$$t_d = t_i - t_{i-1}$$

で求める。そして、求まったａ、t_dの値に応じて、次の４種類の処理のうちのいずれかを実行する。

（１）水平方向への連続移動をしていることを示すフラグr_{flag}がオフで、ａ＜Ａｒかつt_d＜Ｔｒ（傾き及び直前のイベントとの時間間隔がしきい値の範囲内）である場合：マウスによる水平方向への連続移動が始まったと解釈し、フラグr_{flag}をオンにして、マウスポインタの座標（x_i，y_i）を記憶する。

（２）フラグr_{flag}がオフでａ≧Ａｒまたはt_d≧Ｔｒ（傾きまたは直前のイベ

ントとの時間間隔のうち少なくともいずれか一方がしきい値の範囲を越えている）である場合：マウスによる水平方向の連続移動が行われていないと解釈する。

（３）フラグr_{flag}がオンで、ａ＜Ａｒかつt_d＜Ｔｒである場合：マウスによる水平方向への連続移動中であると解釈し、マウスポインタの座標（x_i，y_i）を記憶する。

（４）フラグr_{flag}がオンでａ≧Ａｒまたはt_d≧Ｔｒである場合：マウスによる水平方向への連続移動が終了したと解釈し、フラグr_{flag}をオフにする。記憶したマウスの水平方向への連続移動中におけるマウスポインタの座標から、当該動作の始点と終点のｘ座標、当該動作を行っている間のｙ座標の平均、及び移動の長さｌを算出する。ここで、ｌ＜Ｌであれば、抽出された移動の長さがしきい値Ｌよりも短いため、この動作をなぞり読みとしては検出しない。ｌ≧Ｌであれば、この動作をなぞり読みとして検出する。

以上のようにして、なぞり読み操作が検出されたならば、操作イベント解析部２０は、テキスト抽出部３０に通知し、mousemoveイベントにより得られたなぞり読みの始点及び終点におけるマウスポインタの座標（位置情報）をテキスト抽出部３０に送る。

【００４７】

テキスト抽出部３０は、なぞり読み操作が行われたことを知らせる通知を受け取ると、当該なぞり読みの行われた箇所のテキストを抽出する。ここで、テキストの抽出は、なぞり読みを行ったときにマウスポインタの重なっている行のテキスト及びその一つ上の行のテキストを、文または行の単位でそれぞれ抽出する。マウスポインタの重なっている行及び一つ上の行を抽出する理由は、ユーザはなぞり読みを行っている際に、マウスポインタと同じ行または一つ上の行を読む傾向があるためである。したがって、マウスポインタの重なっている行及び一つ上の行を抽出することにより、ユーザが興味を持って注目したと考えられる情報が抽出対象から外れにくくなる。なお、マウスポインタの重なっている行及び一つ上の行の両方からテキストを抽出するのではなく、いずれか一方からのみ抽出するようにしても良い。

マウスポインタの重なっている行の一つ上の行を認識するためには、マウスポ

インタの位置を基準として、ｙ座標の上の方向に順次着目し、検出される文字列が変わったところで一つ上の行に移ったと判断する。具体的には、まず、マウスポインタの存在する場所の文字とそのｍ個前の文字及びｎ個後の文字の合計三文字を記憶する。ここで、ｍ、ｎは２以上の適当な数である。この後、マウスポインタの位置（座標値）からｙ座標の上の方向に数ドットずつ座標を動かして着目してゆき、順次着目する座標の文字とそのｍ個前の文字及びｎ個後の文字を取得する。そして、これらを先に記憶したマウスポインタの存在する場所の文字及びその前後の文字と比較する。比較の結果三つの文字が全て一致した場合は、未だマウスポインタが重なっている行と認識し、それ以外は、一つ上の行であると認識する。

【００４８】

マウスポインタの重なっている行とその一つ上の行とを識別するために、マウスポインタの存在する場所の文字と前後に数個ずつ離れた二つの文字の合計３文字を用いる理由について説明する。まず、一文字のみで認識しようとすると、偶然マウスポインタのある場所の文字とその上に位置する文字とが一致してしまう場合がある。そこで、信頼性を高めるために複数の文字を用いて行の認識を行う。また、マウスポインタのある場所の文字から数個離れた文字を用いるのは、マウスポインタの重なっている文字が含まれる単語の上に偶然同一の単語が存在する場合は、マウスポインタの位置の文字を含む数文字が上下の行で同一となるため、これを排除する必要があるためである。さらに、マウスポインタのある場所の文字の前と後に位置する文字をそれぞれ用いるのは、マウスポインタのある場所の文字がページの先頭や末尾である場合、前方のみまたは後方のみから文字を取ると、マウスポインタのある文字から数個離れた文字の位置が当該ページから外れてしまう場合があるが、そのような場合でもマウスポインタのある場所の文字の前後から比較用の文字を取っておけば、行の識別を行うことが可能となるためである。

マウスポインタが重なっている行の一つ上の行を認識する手順を、図７を参照して説明する。図７において、まず、マウスポインタの重なっている行の三つの文字（”j”,”r”,”e”）を記憶する。その後、着目する座標（後述するTextRangeオ

ブジェクトの選択範囲）を、マウスポインタのある座標からｙ座標の上方向に数ドットずつ動かし、その座標にある行の三つの文字（"i","r","o"）を取得する。ここでは、マウスポインタのある文字は"r"と"r"で同一であるが、"j"と"i"、及び"e"と"o"の二組が異なるため、一つ上の行であると認識できる。

【００４９】

以上の前提に基づいて、テキスト抽出部３０がなぞり読みの対象となった行及びその一つ上の行からテキストを抽出する処理について説明する。図８は、上述したテキスト抽出部３０によるテキスト抽出処理を、ダイナミックＨＴＭＬを用いて実現する場合のプログラムを説明する図である。

ここでは、なぞり読みを検出した際におけるテキストの抽出のために、getTracedText関数を定義する。getTracedText関数は、操作イベント解析部２０でなぞり読みを検出した後に、マウスポインタの座標を用いてマウスポインタの重なっている行または一つ上の行のテキストを抽出する関数である。引数のｘ、ｙはマウスポインタの存在する座標（ｘ，ｙ）である。また、ｕｐは抽出する行を示し、ｕｐ＝falseの場合はマウスポインタと重なっている行を抽出し、ｕｐ＝trueの場合はマウスポインタと重なっている行の一つ上の行を抽出する。

図８のプログラムリストの３行目において、TextRangeオブジェクトを生成する。また、４行目において、マウスポインタの存在する座標（ｘ，ｙ）にある文字に、TextRangeオブジェクトの選択範囲を移動する。

プログラムリストの５～２５行の処理は、マウスポインタと重なっている行の一つ上の行を認識する処理である。まず、７～１１行において、マウスポインタの重なっている行にある三つの文字（centerchar1，rightchar1，leftchar1）を取得する。これは、マウスポインタの位置にある文字（centerchar1）とその後方ＣＭＯＶＥ文字先の文字（rightchar1）及び前方ＣＭＯＶＥ文字先の文字（leftchar1）である。

次に、１２～２４行において、マウスポインタの存在する位置からＰＭＯＶＥポイントずつｙ座標の上方向に移動し、その位置での文字とその前後ＣＭＯＶＥ文字ずつ先の２文字の合計３文字（centerchar2，rightchar2，leftchar2）を取得する。そして、７～１１行で取得した（centerchar1，rightchar1，leftchar1

）と比較する。この中で一つでも異なるものがあれば、一つ上の行と認識する。

この後、２６行目で、文または行の単位までTextRangeオブジェクトの選択範囲を拡張し、２７行目で、その行のテキストを抽出する。マウスが重なっている行またはその一つ上の行において、文または行の単位でテキストを抽出する方法は、リンクのポインティング操作において説明した方法と同様である。

【００５０】

次に、特定の操作として縦方向なぞり読みが行われた場合について説明する。

縦方向なぞり読みは、読んでいるテキストの行をマウスポインタで指しながら、マウスポインタを行と直交する方向に少しずつ動かす読み方で、移動距離の短い、非常にゆっくりとしたマウスの動きとなる。縦方向なぞり読みの抽出は、マウスの移動イベントで得られるマウスポインタの座標（ｘ，ｙ）とイベント発生時刻を利用して行う。縦方向なぞり読みを検出する方法は、いくつか考えられるが、本実施の形態では以下の方法を用いる。

まず、マウスの動きに途切れのない垂直方向への連続移動を検出する。そして、垂直方向への連続移動の長さがしきい値以上の場合に縦方向なぞり読みとして検出する。これは、偶然マウスが垂直方向の直線移動した動作を排除するためである。なお、垂直方向への連続移動は、マウスの移動イベントが発生するたびに、次の条件を判断することにより検出できる。

第１に、最後のイベントにおけるマウスポインタの座標とその直前のイベントにおけるマウスポインタの座標の変位とから、マウスがウィンドウ上で垂直方向の移動をしているかどうかを判断する（縦方向なぞり読みのマウスの動きは非常に遅くかつ短いため、傾きではなくマウスポインタの座標の変位に基づいて判断する）。

第２に、最後のイベント発生時刻とその直前のイベント発生時刻との差からマウスの動きが途切れていないかどうかを判断する。

以上の条件を満足して、マウスが垂直方向の移動をし、かつ動きが途切れていないと判断された場合、マウスは垂直方向への連続移動をしていると認識する。そして、この二つの条件のどちらかが満たされなくなった場合、垂直方向への連続移動が終了したと判断する。

【００５１】

以上の前提に基づいて、操作イベント解析部２０が縦方向なぞり読みの操作を検出する処理について説明する。以下の説明において、パラメータＸｒ、Ｙｒは、マウスの移動方向を垂直方向とみなすかどうかを決定するためのマウスの移動における変位のしきい値である。パラメータＴｒは、マウスの動作が連続移動であるかどうかを決定するための停止時間に関するしきい値である。また、パラメータＬは、検出された垂直方向への連続移動を縦方向なぞり読みと判断するための移動の長さに関するしきい値である。なお、座標は直交するｘ－ｙ座標にて表され、ｘ方向が表示画面の水平方向（すなわち、行と平行な方向）、ｙ方向が表示画面の垂直方向（すなわち、行と直交する方向）とする。

操作イベント解析部２０は、mousemoveイベントが発生するたびに、当該マウスポインタの座標（x_i，y_i）とその直前に発生したmousemoveイベントにおけるマウスポインタの座標（x_{i-1}，y_{i-1}）との差（$x_i - x_{i-1}$，$y_i - y_{i-1}$）を求める。そして、$0 < y_i - y_{i-1} < Yr$であれば、次に、ｘ方向の差の絶対値ｄを下式

$$d = | x_i - x_{i-1} |$$

で求める。また、最後のイベント発生時刻t_iとその直前のイベント発生時刻t_{i-1}との間の時間間隔t_dを下式

$$t_d = t_i - t_{i-1}$$

で求める。そして、求まったｄ、t_dの値に応じて、次の４種類の処理のうちのいずれかを実行する。

（１）垂直方向への連続移動をしていることを示すフラグr_{flag}がオフで、$d < Xr$かつ$t_d < Tr$（ｘ方向の変位及び直前のイベントとの時間間隔がしきい値の範囲内）である場合：マウスによる垂直方向への連続移動が始まったと解釈し、フラグr_{flag}をオンにして、マウスポインタの座標（x_i，y_i）を記憶する。

（２）フラグr_{flag}がオフで$d \geqq Xr$または$t_d \geqq Tr$（ｘ方向の変位または直前のイベントとの時間間隔のうち少なくともいずれか一方がしきい値の範囲を越えている）である場合：マウスによる垂直方向の連続移動が行われていないと解釈する。

（３）フラグr_{flag}がオンで、ｄ＜Ｘｒかつt_d＜Ｔｒである場合：マウスによる垂直方向への連続移動中であると解釈し、マウスポインタの座標（x_i，y_i）を記憶する。

（４）フラグr_{flag}がオンでｄ≧Ｘｒまたはt_d≧Ｔｒである場合：マウスによる垂直方向への連続移動が終了したと解釈し、フラグr_{flag}をオフにする。記憶したマウスの垂直方向への連続移動中におけるマウスポインタの座標から、当該動作の始点と終点のｙ座標、当該動作を行っている間のｘ座標の平均、及び移動の長さｌを算出する。ここで、ｌ＜Ｌであれば、抽出された移動の長さがしきい値Ｌよりも短いため、この動作を縦方向なぞり読みとしては検出しない。ｌ≧Ｌであれば、この動作を縦方向なぞり読みとして検出する。

以上のようにして、縦方向なぞり読み操作が検出されたならば、操作イベント解析部２０は、テキスト抽出部３０に通知し、mousemoveイベントにより得られたなぞり読みの始点及び終点におけるマウスポインタの座標（位置情報）をテキスト抽出部３０に送る。

【００５２】

テキスト抽出部３０は、縦方向なぞり読み操作が行われたことを知らせる通知を受け取ると、当該なぞり読みの行われた箇所のテキストを抽出する。ここで、テキストの抽出は、縦方向なぞり読みを行ったときにマウスポインタの重なっている行のテキスト及びその一つ上の行のテキストを、文または行の単位でそれぞれ抽出する。また、マウスポインタの重なっている行または一つ上の行のいずれか一方からのみ抽出するようにしても良い。

テキスト抽出部３０による具体的なテキスト抽出の処理は、上述したなぞり読みの場合と同様であるため、説明を省略する。

【００５３】

なお、以上の説明において、マウスポインタをテキストの行に沿って横方向に動かす動作をなぞり読みと呼び、テキスト中の読んでいる行をマウスポインタで指し、マウスポインタを行に直交する方向に少しずつ動かす動作を縦方向なぞり読みと呼んだが、これはテキストが横書きであることを前提としたためである。テキストが縦書きである場合は、行に沿った縦方向の動作がなぞり読みであり、

行と直交する横方向の動作が縦方向なぞり読みに相当する。

【００５４】

以上説明した本実施の形態における情報抽出システムは、インターネットなどのネットワークに接続し、ウェブブラウザを搭載した情報処理装置において機能する。すなわち、当該情報処理装置がウェブサーバから受信したウェブコンテンツをウェブブラウザにて表示し、表示された当該ウェブコンテンツの内容をユーザが閲覧する際に、無意識に行う上記の各操作を操作イベントとして検出し、そのような操作の対象となったテキストを抽出する。

ここで、情報処理装置に本実施の形態における情報抽出システムの機能を提供する手段として、種々の形態が考えられる。以下、代表的な提供形態について図９乃至図１２を参照して説明する。

【００５５】

図９に示す提供形態は、操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０をJavaScript等のスクリプト言語で記述し、ウェブサーバ１００に格納されるウェブコンテンツ１０１に予め埋め込んでおく形態である。これにより、情報処理装置１１０がウェブサーバ１００からウェブコンテンツ１０１を受信すると、ウェブブラウザ１１１は、ウェブコンテンツ１０１に埋め込まれているスクリプト１０２に基づいて、上述した操作イベントを検出する処理、操作イベント列を解析して文字列の選択、リンクのポインティング、なぞり読み等の上記特定の操作を検出する処理、及び当該操作の対象である文字列を抽出する処理を実行する。抽出されたテキストは、ウェブサーバ１００に返送される。抽出されたテキストをウェブサーバ１００に返送する機能は、操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０と同様に、スクリプトとしてウェブコンテンツ１０１に埋め込むことによって提供しても良いし、Ｊａｖａアプレット等の形でウェブコンテンツ１０１と一緒に情報処理装置１１０に配信しても良い。

こうして得られたテキストは、ユーザがウェブコンテンツ１０１を閲覧する際に、興味を持った情報であると考えられるため、ウェブサーバ１００において、ウェブ視聴率の調査や、検索エンジンにおける検索条件の絞り込み等、種々のサ

ービスに利用することができる。

【００５６】

図１０に示す提供形態は、ウェブサーバ１００が、ウェブコンテンツ１０１に対して操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０をJavaScript等のスクリプト言語で書き込む書き込み処理部１２０を備える形態である。この形態では、情報処理装置１１０からウェブコンテンツ１０１へのアクセス要求が出されると、ウェブサーバ１００では、まず書き込み処理部１２０により、当該ウェブコンテンツ１０１に対して操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０の機能を実現するスクリプトが書き込まれる。その上で、当該ウェブコンテンツ１０１が情報処理装置１１０に送信される。

情報処理装置１１０のウェブブラウザ１１１は、受信したウェブコンテンツ１０１に埋め込まれているスクリプトに基づいて、上述した操作イベントを検出する処理、操作イベント列を解析して文字列の選択、リンクのポインティング、なぞり読み等の上記特定の操作を検出する処理、及び当該操作の対象である文字列を抽出する処理を実行する。抽出されたテキストは、ウェブサーバ１００に返送される。抽出されたテキストをウェブサーバ１００に返送する機能は、操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０と同様に、スクリプトとしてウェブコンテンツ１０１に埋め込むことによって提供しても良いし、Ｊａｖａアプレット等の形でウェブコンテンツ１０１と一緒に情報処理装置１１０に配信しても良い。

こうして得られたテキストは、ユーザがウェブコンテンツ１０１を閲覧する際に、興味を持った情報であると考えられるため、ウェブサーバ１００において、ウェブ視聴率の調査や、検索エンジンにおける検索条件の絞り込み等、種々のサービスに利用することができる。

【００５７】

図１１に示す提供形態は、ウェブサーバ１００と情報処理装置１１０との間にプロキシサーバ１３０を介在させ、当該プロキシサーバ１３０において、ウェブコンテンツ１０１に対して操作イベント検出部１０、操作イベント解析部２０及

びテキスト抽出部３０をJavaScript等のスクリプト言語で書き込む形態である。この形態では、情報処理装置１１０からウェブコンテンツ１０１へのアクセス要求が出され、ウェブサーバ１００から情報処理装置１１０へウェブコンテンツ１０１が送られると、まずプロキシサーバ１３０が当該ウェブコンテンツ１０１を受信する。そして、当該ウェブコンテンツ１０１に対して操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０の機能を実現するスクリプトを書き込み、その上で、当該ウェブコンテンツ１０１を情報処理装置１１０に送信する。

情報処理装置１１０のウェブブラウザ１１１は、受信したウェブコンテンツ１０１に埋め込まれているスクリプトに基づいて、上述した操作イベントを検出する処理、操作イベント列を解析して文字列の選択、リンクのポインティング、なぞり読み等の上記特定の操作を検出する処理、及び当該操作の対象である文字列を抽出する処理を実行する。抽出されたテキストは、プロキシサーバ１３０に送信される。抽出されたテキストをプロキシサーバ１３０に送信する機能は、操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０と同様に、スクリプトとしてウェブコンテンツ１０１に埋め込むことによって提供しても良いし、Ｊａｖａアプレット等の形でウェブコンテンツ１０１と一緒に情報処理装置１１０に配信しても良い。

こうして得られたテキストは、ユーザがウェブコンテンツ１０１を閲覧する際に、興味を持った情報であると考えられるため、プロキシサーバ１３０において、ウェブ視聴率の調査や、検索エンジンにおける検索条件の絞り込み等、種々のサービスに利用することができる。

【００５８】

また、図１１の提供形態の変形例として、プロキシサーバ１３０はウェブコンテンツ１０１に対して操作イベント解析部２０及びテキスト抽出部３０の機能を実現するスクリプトの埋め込みを行わず、情報処理装置１１０において操作イベントの検出のみを実行させるようにしても良い。この場合、プロキシサーバ１３０に操作イベント解析部２０及びテキスト抽出部３０が設けられる。情報処理装置１１０において検出された操作イベントは、プロキシサーバ１３０に送られる

。そして、プロキシサーバ１３０において、操作イベント列を解析して文字列の選択、リンクのポインティング、なぞり読み等の上記特定の操作を検出する処理、及び当該操作の対象である文字列を抽出する処理が実行される。

情報処理装置１１０において検出された操作イベントをプロキシサーバ１３０に転送するには、プロキシサーバ１３０がウェブコンテンツ１０１を情報処理装置１１０に送る際に、当該ウェブコンテンツ１０１に操作イベントを返送するためのスクリプトを埋め込んでも良いし、プロキシサーバ１３０から操作イベントの送信を情報処理装置１１０に対して要求して送信させるようにしても良い。また、プロキシサーバ１３０は、ウェブサーバ１００から受信したウェブコンテンツ１０１を保持しておき、当該保持されたウェブコンテンツ１０１からテキスト抽出部３０がテキストを抽出するようにしても良いし、該当するウェブコンテンツ１０１を情報処理装置１１０から転送するようにしても良い。

【００５９】

図１２に示す提供形態は、情報処理装置１１０がインターネットに接続した際に最初にアクセスするポータルサイト１４０が、ウェブコンテンツ１０１に対して操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０をJavaScript等のスクリプト言語で書き込むローカルプロキシを実現するプログラムファイル１５０を情報処理装置１１０に送信する形態である。この形態では、情報処理装置１１０がポータルサイト１４０にアクセスすると、ポータルサイト１４０の記憶部１４１に格納されたプログラムファイル１５０が送受信部１４２を介して情報処理装置１１０に送信される。プログラムファイル１５０はＪａｖａアプレット等として作成される。ポータルサイト１４０から情報処理装置１１０へ送信されたプログラムファイル１５０は、情報処理装置１１０においてローカルプロキシ１６０として動作する。このローカルプロキシ１６０は、情報処理装置１１０がウェブサーバ１００から受信したウェブコンテンツ１０１に対して操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０の機能を実現するスクリプトを書き込み、ウェブブラウザ１１１に渡す。

ウェブブラウザ１１１は、ローカルプロキシ１６０から受け取ったウェブコンテンツ１０１に埋め込まれているスクリプトに基づいて、上述した操作イベント

を検出する処理、操作イベント列を解析して文字列の選択、リンクのポインティング、なぞり読み等の上記特定の操作を検出する処理、及び当該操作の対象である文字列を抽出する処理を実行する。抽出されたテキストは、ポータルサイト１４０に送信される。抽出されたテキストをプロキシサーバ１３０に送信する機能は、操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０と同様に、スクリプトとしてウェブコンテンツ１０１に埋め込むことによって提供しても良いし、ローカルプロキシ１６０の機能として提供しても良い。また、ポータルサイト１４０の送受信部１４２において、抽出されたテキストを送信するように情報処理装置１１０に要求を出すことにより、抽出されたテキストの収集を行っても良い。

こうして得られたテキストは、ユーザがウェブコンテンツ１０１を閲覧する際に、興味を持った情報であると考えられるため、ポータルサイト１４０において、ウェブ視聴率の調査や、検索エンジンにおける検索条件の絞り込み等、種々のサービスに利用することができる。

【００６０】

図１３は、このようにして得られたテキストを用いて、検索エンジンにおけるキーワードベクトル（選択されたキーワードとその重要度を示す重み）を生成する場合について、本実施の形態と従来技術とを比較する図である。従来技術の手法では、ページ全体に含まれるキーワードにＴＦ・ＩＤＦ式等の手法によりキーワードに重みを付けて、重要なキーワードを抽出している。これに対し、本実施の形態では、ページ中でユーザが行った操作の対象となったテキストに対して、キーワードの重み付け処理を行う。キーワードの重み付けに関しては、ＴＦ・ＩＤＦ式におけるＩＤＦ式のような既存の方法を用いることができる。本実施の形態により得られたテキストに基づいて生成されたキーワードベクトルは、単独でウェブ視聴率の調査や、検索エンジンにおける検索条件の絞り込み等のサービスに用いることもできるが、図１３に示すように、従来技術において生成されたキーワードベクトルと組み合わせて利用することもできる。

【００６１】

なお、以上の四種類の提供形態において、抽出されたテキストの送信先は、上

記の各送信先に限定されない。当該テキストを利用できる種々の利用者に対して送信することができる。例えば、図９に示した形態において、スクリプト１０２を埋め込んだウェブコンテンツ１０１の作成者に対して、抽出されたテキストを送るようにしても良い。また、図１１や図１２に示した形態において、当該抽出されたテキストを用いたサービスを行う、プロキシサーバ１３０やポータルサイト１４０とは別に設けられたサーバに転送するようにしても良い。

【００６２】

また、本実施の形態では、ウェブコンテンツを対象として、ユーザが行った操作に基づいてテキストの抽出を行ったが、他の任意の形式の文書データを対象としてテキストの抽出を行うようにしても良い。この場合、マウスの動作を監視する領域としては、ウェブブラウザによるウェブコンテンツの表示画面以外に、コンピュータのディスプレイ装置における表示画面全体や、アプリケーションプログラムによって表示されるウィンドウの領域内など、種々の範囲に設定することができる。

【００６３】

さらにまた、テキスト以外の画像などのオブジェクトに対してユーザが行った操作に基づいて、当該操作の対象となったオブジェクトに関する情報を抽出することもできる。この場合、ユーザが興味を持って無意識的に行った動作として定義される動作としては、テキスト選択と同様にして行われたオブジェクトの選択操作や、リンクのポインティング、クリックなどが挙げられる。

【００６４】

さらに、マウスや他のポインティングデバイス以外の入力手段を用いて、ユーザが興味を持って無意識的に行った動作を定義することも可能である。例えば、カーソルキーなどを用いたキー操作や、ユーザが表示されたテキストを読み上げた場合の音声の入力や、ユーザの視線の動き等から特定の動作を定義することができる。

【００６５】

【発明の効果】

以上説明したように、本発明によれば、ユーザによる明示的な入力を必要とせ

ず、かつウェブコンテンツにおいてユーザが興味を持った箇所に関する詳細な情報を取得することができる。

また、ユーザによるウェブブラウザ上での操作を含む詳細な操作を抽出して、ユーザの抱く興味の傾向を示す情報として利用することができる。

【図面の簡単な説明】

【図１】　本実施の形態における情報抽出システムの全体構成を説明するための図である。

【図２】　本実施の形態における操作イベント検出部１０、操作イベント解析部２０及びテキスト抽出部３０の働きを概念的に説明する図である。

【図３】　テキスト選択が行われた場合のテキスト抽出部３０によるテキスト抽出処理を、ダイナミックＨＴＭＬを用いて実現する場合のプログラムを説明する図である。

【図４】　テキスト選択が行われた場合のテキスト抽出の課程を説明する図である。

【図５】　リンクのポインティングが行われた場合のテキスト抽出部３０によるテキスト抽出処理を、ダイナミックＨＴＭＬを用いて実現する場合のプログラムを説明する図である。

【図６】　リンクのポインティングが行われた場合のテキスト抽出の課程を説明する図である。

【図７】　なぞり読みにおいて、マウスポインタが重なっている行の一つ上の行を認識する処理を説明する図である。

【図８】　なぞり読みが行われた場合のテキスト抽出の課程を説明する図である。

【図９】　本実施の形態における情報抽出システムを提供する形態の例を説明する図である。

【図１０】　本実施の形態における情報抽出システムを提供する形態の他の例を説明する図である。

【図１１】　本実施の形態における情報抽出システムを提供する形態の更に他の例を説明する図である。

【図１２】　本実施の形態における情報抽出システムを提供する形態の更に他の例を説明する図である。

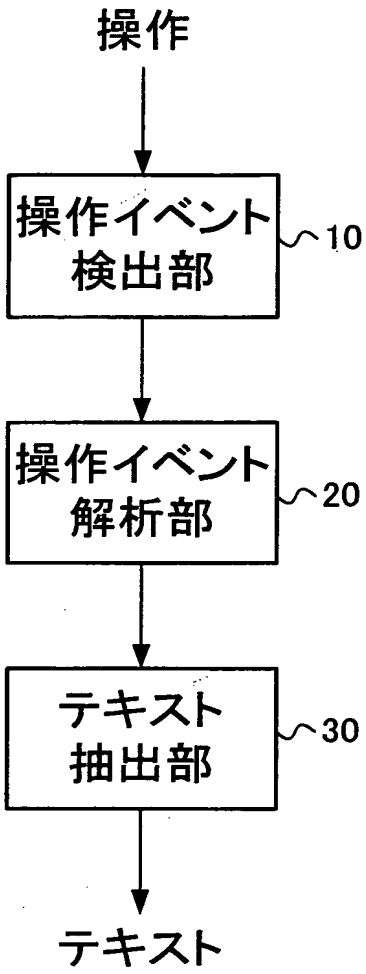
【図１３】　本実施の形態により抽出されたテキストを検索エンジンにおけるキーワードベクトルの生成に用いた例を従来技術と比較する図である。

【符号の説明】

１０…操作イベント検出部、２０…操作イベント解析部、３０…テキスト抽出部、１００…ウェブサーバ、１０１…ウェブコンテンツ、１０２…スクリプト、１１０…情報処理装置、１１１…ウェブブラウザ、１２０…書き込み処理部、１３０…プロキシサーバ、１４０…ポータルサイト、１４１…記憶部、１４２…送受信部、１５０…プログラムファイル、１６０…ローカルプロキシ、２００…ウェブコンテンツ、２０１…テキスト、４０１…selectionオブジェクト、４０２…TextRangeオブジェクト、４０３…テキスト、６０１、６０２、６０３…TextRangeオブジェクト、６０４…テキスト

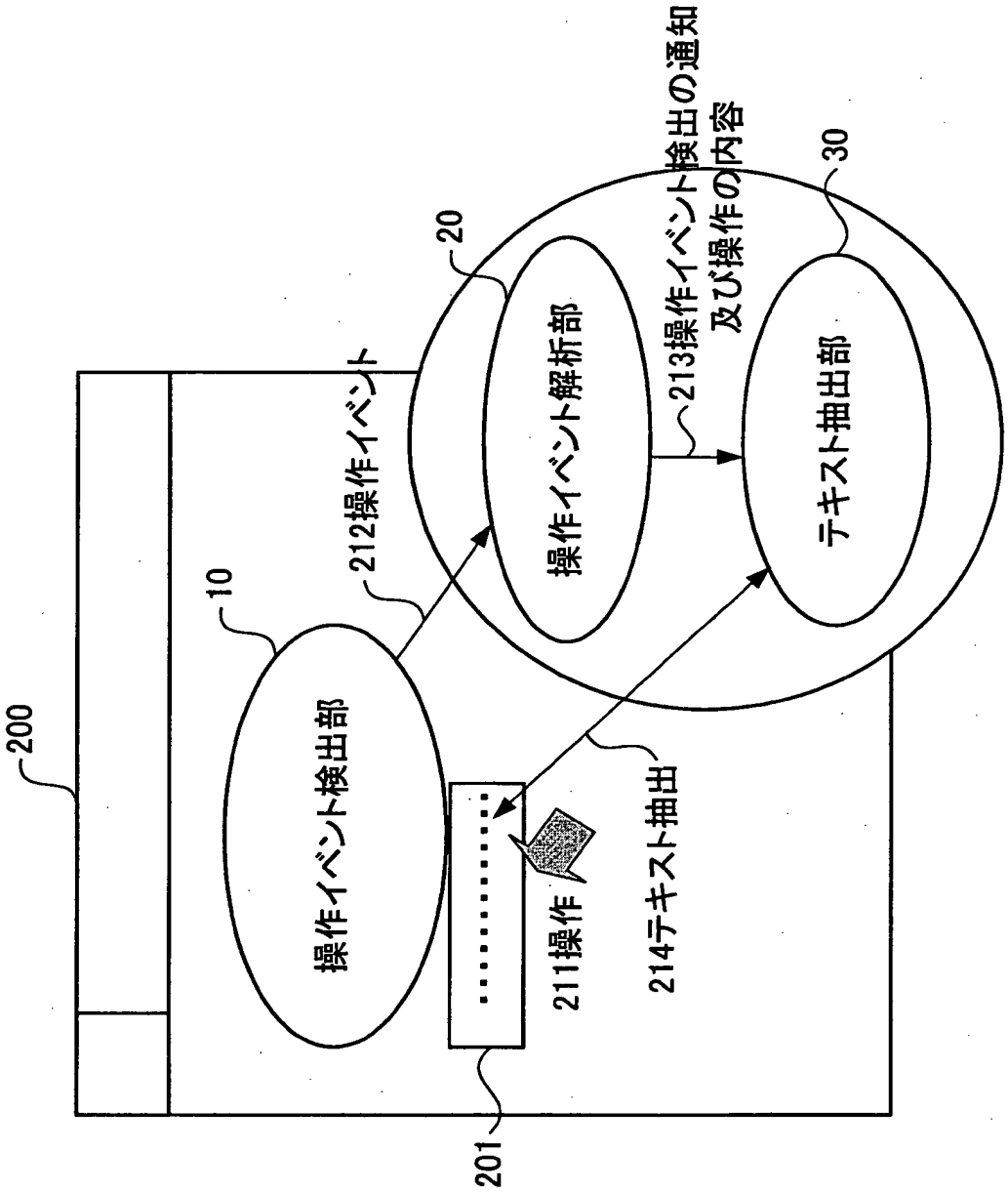
【書類名】　　図面

【図１】



出証特２０００－３０５１３０７

【図２】

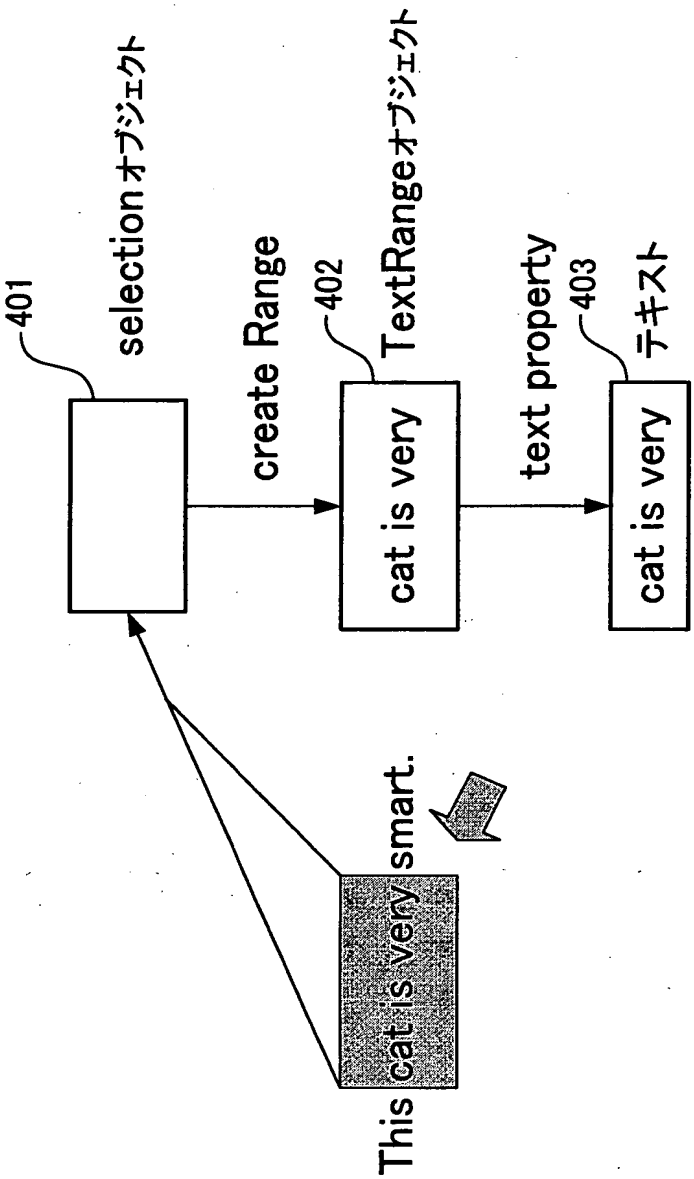

200
10
操作イベント検出部
212操作イベント
20
操作イベント解析部
213操作イベント検出の通知
及び操作の内容
30
テキスト抽出部
211操作
214テキスト抽出
201

【図３】

プログラムリスト

```
1:function getSelectedText(sl)
2:{
3:      tr=sl.createRange();
4:      return tr.text
5:}
```

【図４】

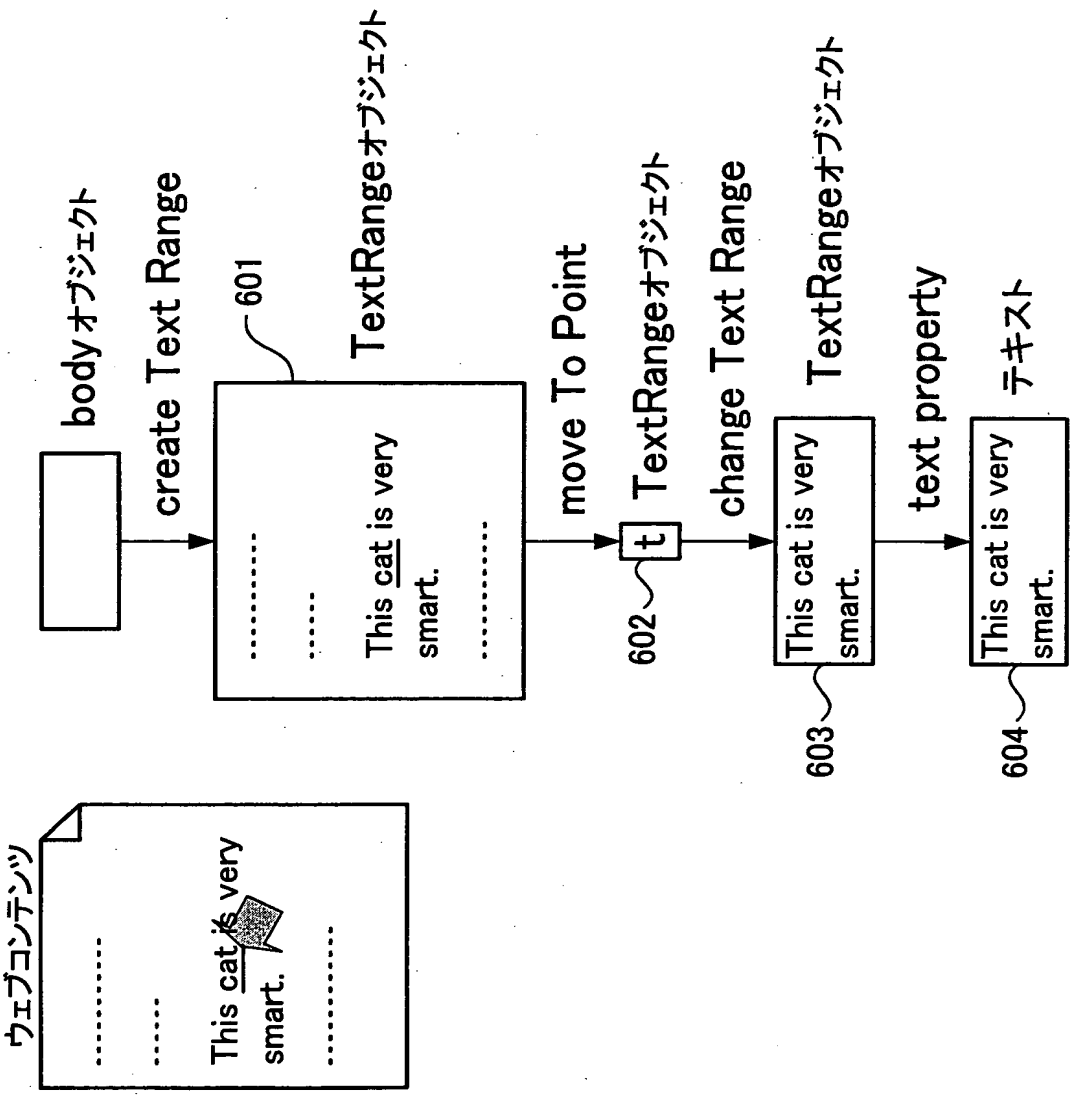

401
selectionオブジェクト
create Range
402
TextRangeオブジェクト
text property
403
テキスト
cat is very
cat is very
This cat is very smart.

【図５】

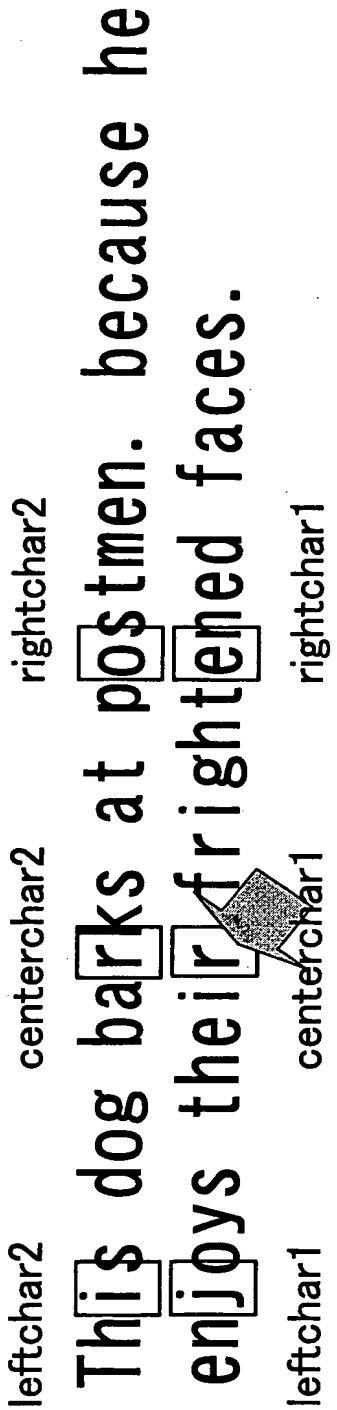
プログラムリスト

```
1:function getLinkTagText(anchor)
2:{
3:      return anchor.innerText
4:}
5:
6:function getLinkText(x, y)
7:{
8:      tr=document.body.createTextRange();
9:      tr.moveToPoint(x, y);
10:     changeTaxtRange(tr);
11:     return tr.text;
12:}
```

【図６】


bodyオブジェクト
create Text Range
601
TextRangeオブジェクト
This cat is very smart.
move To Point
602
t
TextRangeオブジェクト
change Text Range
This cat is very smart.
603
TextRangeオブジェクト
text property
This cat is very smart.
604
テキスト
ウェブコンテンツ
This cat is very smart.

【図７】

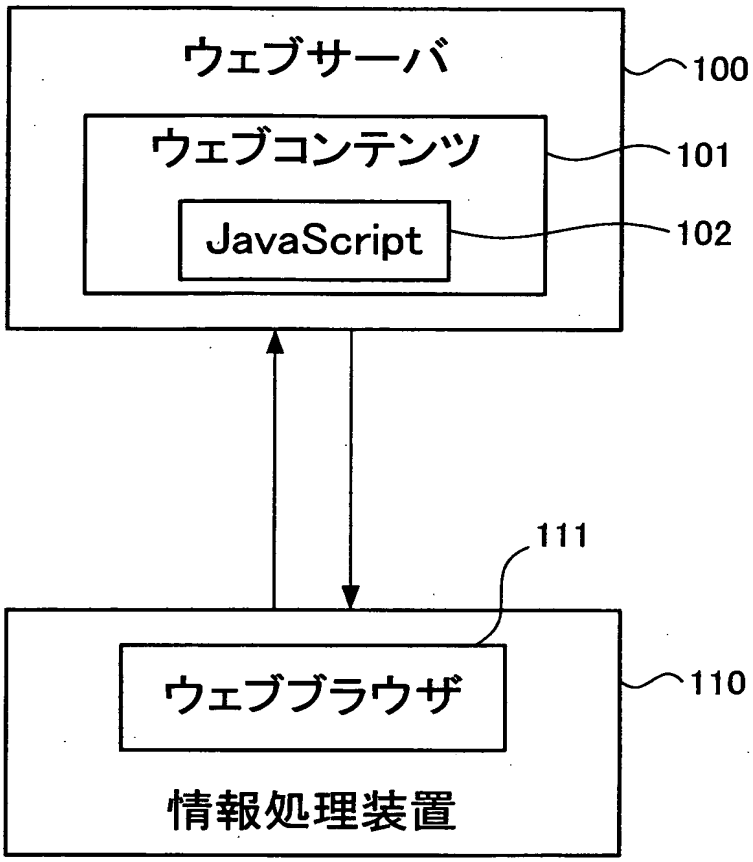

This dog barks at postmen. because he
enjoys their frightened faces.
leftchar2
centerchar2
rightchar2
leftchar1
centerchar1
rightchar1

【図８】

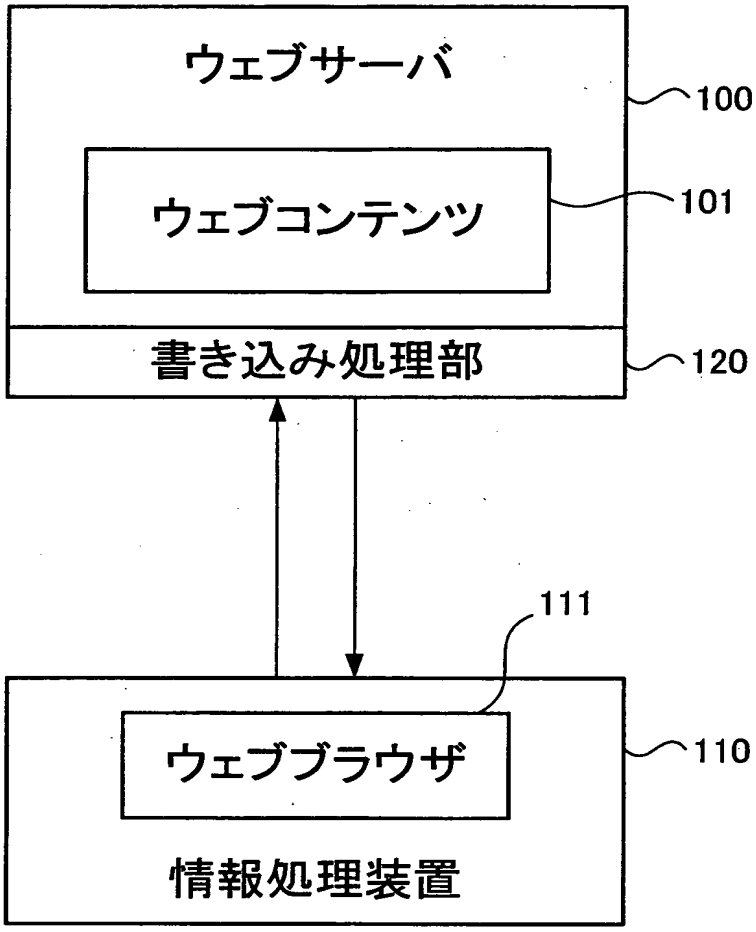
プログラムリスト

```
1:function getTracedText(x, y, up)
2:{
3:      tr=body.createTextRange();
4:      tr.moveToPoint(x, y);
5:      if(up==true)
6:      {
7:              centerchar1=tr.text;
8:              tr.move(CMOVE);
9:              rightchar1=tr.text;
10:             tr.move(-2*CMOVE);
11:             leftchar1=tr.text;
12:             i=PMOVE;
13:             while(i<Um)
14:             {
15:                     tr.moveToPoint(x, y-i);
16:                     centerchar2=tr.text;
17:                     tr.move(CMOVE);
18:                     rightchar2=tr.text;
19:                     tr.move(-2*CMOVE);
20:                     leftchar2=tr.text;
21:                     if(centerchar1!=centerchar2|
                        |rightchar1!=rightchar2|
                        |leftchar1!=leftchar2)
22:                             break;
23:                      i+PMOVE
24:             }
25:             tr.move(CMOVE);
26:     }
27:     changeTextRange(tr);
28:     return.text;
29:}
```

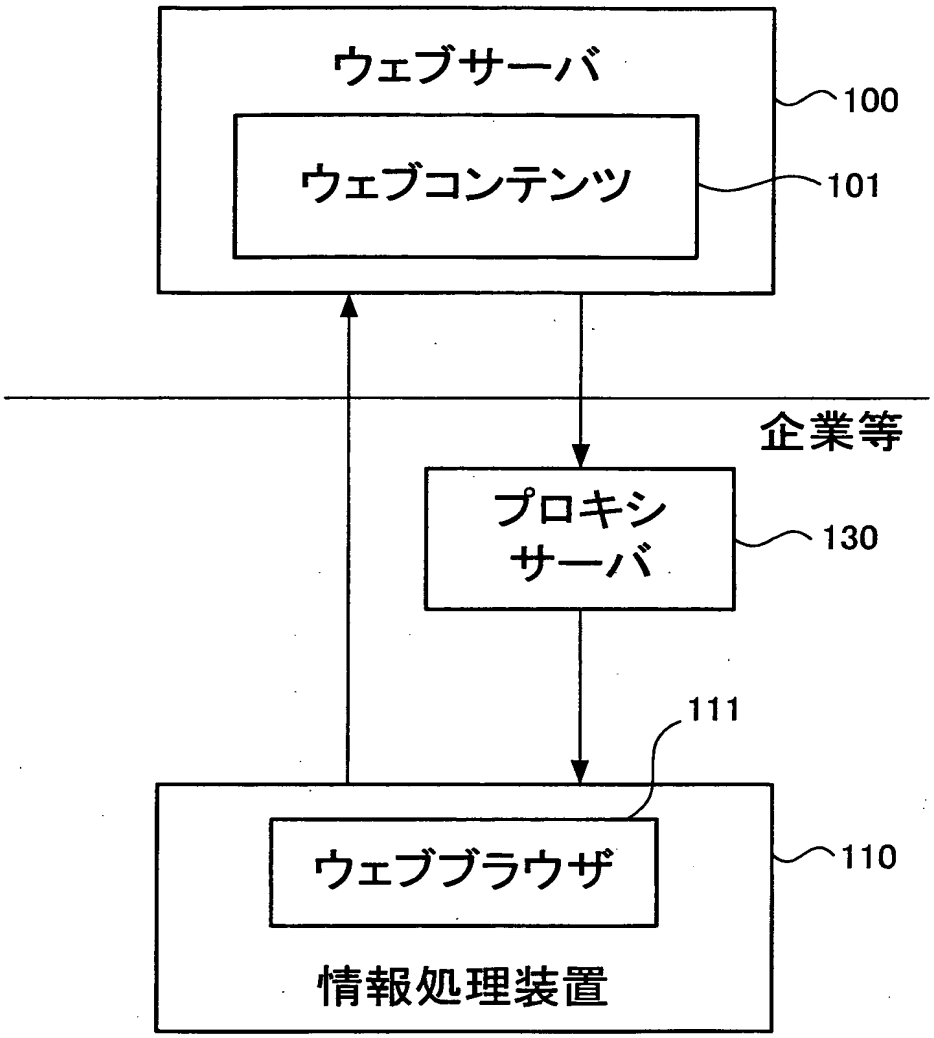
【図９】


ウェブサーバ
100
ウェブコンテンツ
101
JavaScript
102
111
ウェブブラウザ
110
情報処理装置

【図１０】

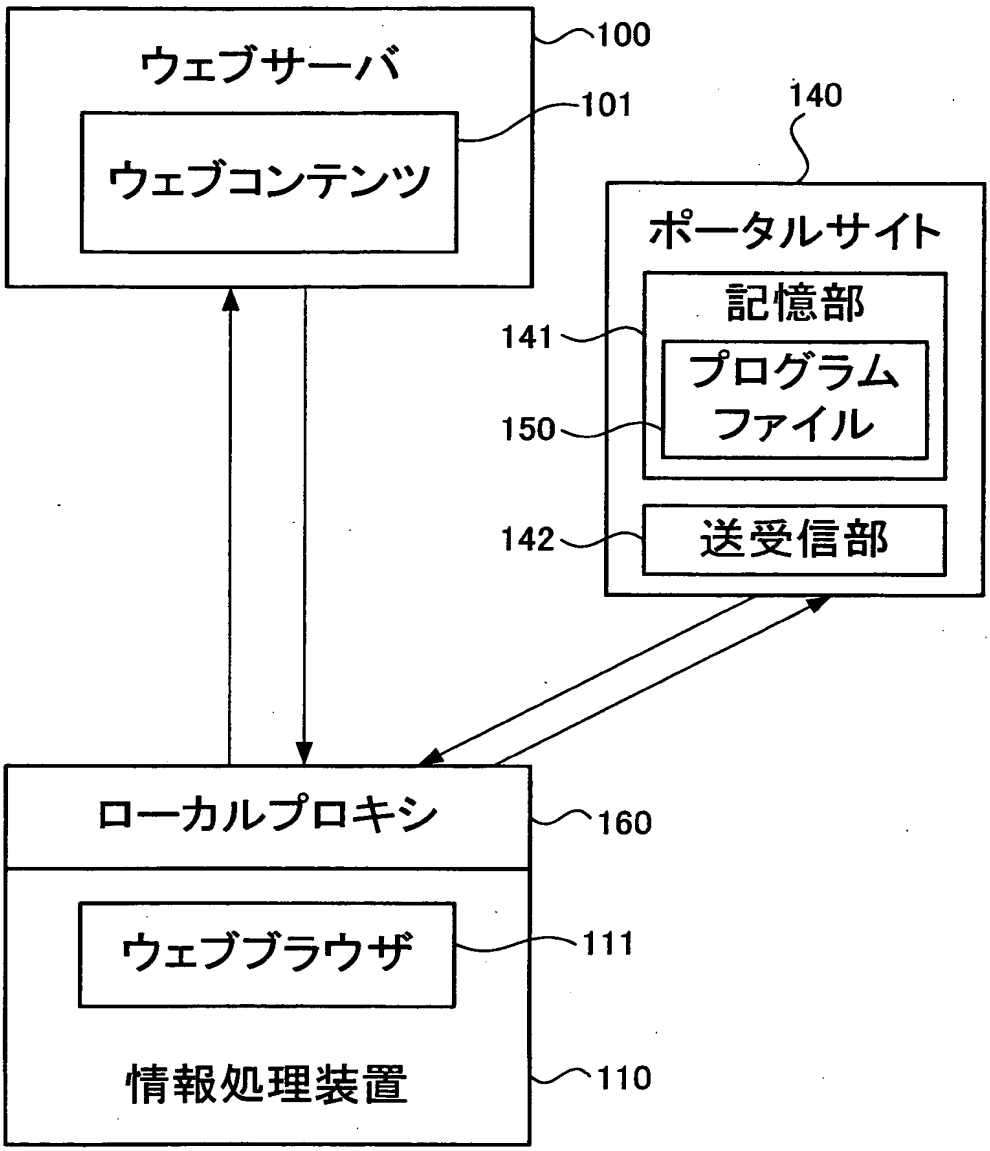


【図１１】



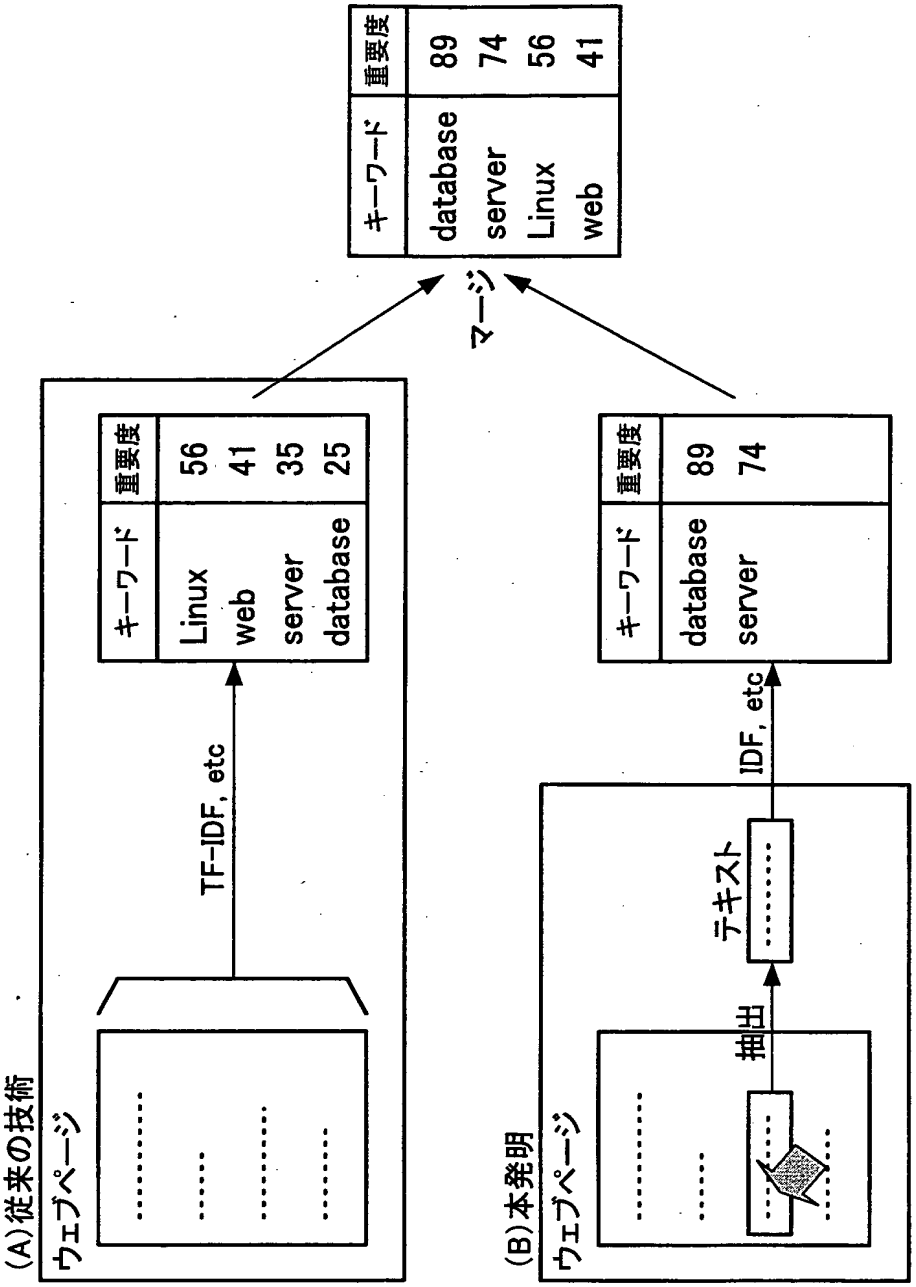
出証特２０００－３０５１３０７

【図１２】



出証特２０００－３０５１３０７

【図１３】


キーワード
重要度
database 89
server 74
Linux 56
web 41
マージ
(A)従来の技術
ウェブページ
TF-IDF, etc
キーワード
重要度
Linux 56
web 41
server 35
database 25
(B)本発明
ウェブページ
抽出
テキスト
IDF, etc
キーワード
重要度
database 89
server 74

【書類名】　　要約書

【要約】

【課題】　ユーザによる明示的な入力を必要とせず、かつウェブコンテンツにおいてユーザが興味を持った箇所に関する詳細な情報を取得できるようにする。

【解決手段】　通信ネットワークにて接続されたサーバとクライアントとを備えた情報抽出システムであって、サーバは、クライアントにおいて閲覧に供されるデータファイルを提供し、クライアントは、通信ネットワークを介してサーバから受信したこのデータファイルの内容を表示するブラウザと、このブラウザにて表示されたデータファイルの内容を閲覧する際にユーザが行った入力操作に基づいて予め定められた特定の操作を検出する操作イベント解析部２０と、この操作イベント解析部２０により検出された、ブラウザの表示画面上における特定の操作の行われた場所に表示されている情報を抽出するテキスト抽出部３０とを備える。

【選択図】　　図１

出証特２０００－３０５１３０７

認定・付加情報

特許出願の番号	平成１１年　特許願　第３７１３４７号
受付番号	５９９０１２７５４０３
書類名	特許願
担当官	塩崎　博子　　１６０６
作成日	平成１２年　２月１５日

＜認定情報・付加情報＞

【特許出願人】

【識別番号】　390009531

【住所又は居所】　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　（番地なし）

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレイション

【代理人】

【識別番号】　100086243

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　坂口　博

【復代理人】　申請人

【識別番号】　100104880

【住所又は居所】　東京都港区赤坂７－１０－９　第４文成ビル２０２セリオ国際特許事務所

【氏名又は名称】　古部　次郎

【選任した代理人】

【識別番号】　100091568

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　市位　嘉宏

【選任した復代理人】

【識別番号】　100100077

【住所又は居所】　東京都港区赤坂７－１０－９　第４文成ビル２０２セリオ国際特許事務所

【氏名又は名称】　大場　充

次頁無

出　願　人　履　歴　情　報

識別番号　　〔３９０００９５３１〕

１．変更年月日　　１９９０年１０月２４日
［変更理由］　新規登録
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　（番地なし）
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレイション

２．変更年月日　　２０００年　５月１６日
［変更理由］　名称変更
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　（番地なし）
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレーション